,,

Universiteit Leiden
The Netherlands

Opleiding Informatica

A Comparative Omics Approach To Understanding Huntington's Disease Network Interactions

Kim van Grondelle

Supervisors:
Katy Wolstencroft & Lu Cao

BACHELOR THESIS

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

27/06/2023

Abstract

This bachelor thesis is about comparing and understanding Huntington's disease network interactions. The comparing was done between humans and two different model organisms commonly used for Huntington's disease. To do this, two protein-protein interaction networks were created, one for mouse and one for rat. These two networks are analyzed and then compared to the already available network for Huntington's disease in humans. Model organisms are used to test and help develop new therapeutics for in this case Huntington's disease. However, these model organisms are not the same as humans. And thus work not completely the same as humans. Trying to identify differences and similarities in the proteins and the interactions between these proteins in the network can help the understanding of Huntington's disease in humans. Which again can help to develop new or improve already existing treatment to this awful disease.

After the comparison can be said that the networks of the two model organisms are almost the same. The proteins in the network are mainly overlapping and also the enriched terms from both networks are the same. However, when comparing the model organisms to humans there is much less overlap between the proteins. Despite the less amount of overlapping proteins there is some overlap between the enriched terms of humans and the model organisms.

Contents

1 Introduction

To begin this thesis, the general idea will be explained. Understanding diseases can be quite hard. Most of the time the symptoms are quite clear and whole lists can be generated. However, understand the processes that cause these symptoms can be hard. Understanding how proteins are connected and interacting in biological systems can help understand the clinical picture of certain diseases. To do this, protein-protein interaction networks will be created. Protein-protein interaction networks are graphs which show interactions between proteins. These interactions can in this case help the understanding of Huntington's disease by identifying (new) protein functions. These functions may be altered or affected by the disease and can be a target for a treatment.

In this thesis different protein-protein interaction networks for Huntington's disease were created, analyzed and compared. To be more specific, networks of proteins according to Huntington's Disease were created for two model organisms that were until recently the only option: *Mus Musculus* and *Rattus Norvegicus*. The networks created for these organisms were analyzed and compared to the human HD networks which was created by Chen Ji Rong Jiang [Jia22] and expanded by Nina Henninger [Hen23]. There have be done three comparisons. With the comparisons possible differences and similarities between mouse, rat and human have were identified.

For the completeness of this thesis, a public GitHub repository was created. All the networks that were created and used in this thesis can be found in the appendices and on GitHub: `https://github.com/Kimvangrondelle/HD_bachelorthesis_s28273249_LeidenUniversiteit`.

1.1 Huntington's Disease

Huntington's Disease is an autosomal-dominant progressive neurodegenerative disorder [Wal07] and is caused by a mutation in the Huntingtin gene. This gene is located on the short arm of the fourth chromosome. The mutation that causes Huntington's disease is an expanded CAG repeat in the protein sequence. A human being without HD has less than 35 CAG repeats. When the count of the CAG repeats is between 35 and 40, the penetrance is reduced, which means that not every one who has this amount of repeats will get HD [HF12]. Over 40 repeats, the penetrance is complete, which means that everyone having 40 or more CAG repeats will get Huntington's disease. The variation in the number of repeats may also influence the age on which the disease manifests [vDvdVRB86, SL11]. In most cases the symptoms start around the age of 40, which can be quite overwhelming, not only because of the diagnosis itself, but also because most of the time those individuals already have children. However, the symptoms can start at every age between infancy and senescence [Wal07].

The CAG repeats are toxic for the function of the geneproducts. The toxicity makes the protein vulnerable for fragmentation which results in dysfunction of neurons and eventually resulting in the death [BDG+15]. Symptoms of Huntington's disease can vary from uncontrollable movements, such as chorea and dystonia and a lack of coordination to cognitive decline and behavioral difficulties [Wal07, HF12]. Nowadays there is unfortunately still no cure thanks in part because the function of the Huntingtin gene is not fully known yet. This makes the treatment only focused on the symptoms instead of the cause of the disease.

Huntingtin is expressed widely through maturing and is dynamically spread within cells. Huntingtin seems to interact with a number of proteins with an effecting function, which makes Huntingtin important as some kind of intermediary to make happen physiological processes [SL11]. The non-mutated variant of Huntingtin is among other things involved in chemical signaling, regulation of transcription and preventing the cell from apoptosis.

Using PPINs can help understand the way in which the proteins are connected. Looking at HD in different organisms which may be used as a model organism, and see how these proteins are connected and may differ or agree from humans, can help the development of new treatments, or improve the already existing treatments for those who suffer from Huntington's disease.

1.2 Existing networks

In this research the research "Extending Consensus Knowledge In Huntington's Disease Protein Interaction Networks" [Hen23] was extended. Whereas [Hen23] looked only at human data, this research was focus on data from both rats and mice. The networks made by [Hen23] were used to compare the networks created for both mice and rats, looking for both similarities and differences between the different organisms.

1.3 Model organisms

Model organisms are organisms used to serve as a model for, for example, disease course research. Researchers as it were, apply a certain disease to those kind of animals. With the use of model organisms, many therapeutic strategies can be tested, and can be seen what the effect is on the disease condition of these animals.

When using model organisms, it has to be taken into account that even if there is a high homology between certain genes for humans and the model organisms, there are always differences between them. These differences lead to other proteins which can cause different reactions to therapeutic strategies [EBPH09]. Identifying those differences is therefore of interest when using model organisms to test a therapeutic strategy helping to cure a disease.

A lot of organisms can be used as a model organism. Commonly used is the *Mus musculus* (house mouse). In this thesis there will be looked at two model organisms in Huntington's disease. Both the *Mus musculus* (house mouse) and the *Rattus Norvegicus* (Norwegian rat) will be investigated. In Figure 1 a picture of both the mouse and rat are shown to prevent some misunderstanding of the appearance of these animals.

1.3.1 Sequence similarity

To calculate the similarity between the HTT genes in the different organisms, UNIPROT BLAST was used. The rat HTT gene has UniprotKB id: p51111. To start the search this UniprotKB ID was used as a search query, for the parameters the default values were chosen. The result showed a 96.9% sequence similarity between the HTT gene in rat and mouse, and a 90.1 % sequence similarity

Figure 1: Picture to show the difference between a mouse and a rat

between the HTT gene in rat and human. Using the mouse UniprotKB ID (p42859) resulted in a sequence similarity of 90.5% between the HTT gene in mouse and human. When aligning the three sequences stood out that the sequences differ the most on the first hundred amino-acids which can be seen in Figure 2. Also can be seen that the human, mouse and rat sequence contain a poly-Q repeat. However, the human one is the most extensive.

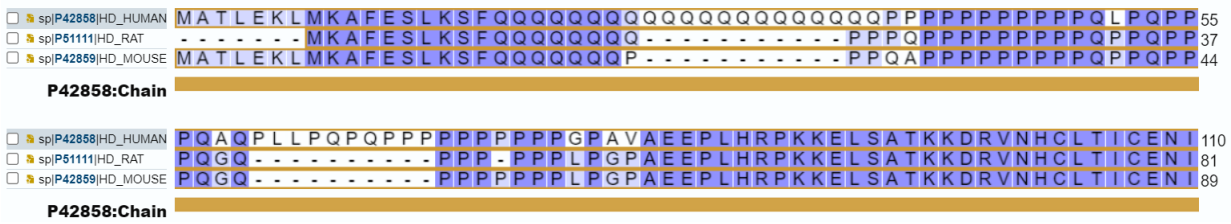

Figure 2: First part of the aligning of the HTT gene sequence of human, mouse and rat. Picture shows the region of the sequences that differs most between the three organisms. Also in the human sequence can be seen that there is an extensive poly-Q repeat.

1.4 Research question

Until recently only mouse and rat were available as model organisms. With this research was investigated whether or not there are common core network features across the three organisms. The research question that was tried to be answered in this thesis is:

"Looking at protein interaction networks according to Huntington's disease, what are the major differences and similarities and are there any common core network features across human, mouse and rat?"

1.5 Thesis overview

To end the introduction of this thesis an overview of the whole thesis was given. This chapter contains the introduction; Section 2 includes the background information of the thesis; Section 3

discusses method used; Section 4 shows the results of the thesis; Section 5 describes the similarities and differences found between the two model organisms and humans. To conclude, there is Section 6 in which the most important results will be discussed together with possible future work.
This bachelor thesis is supervised by Katy Wolstencroft and written for the bachelor Bioinformatics at LIACS.

2 Background

2.1 Definitions

To start the background section some of abbreviations and terms used in this thesis are given in Table 1 together with their corresponding definition.

Table 1: Table with used terms and definitions.

Term	Definition
PPIN	Protein Protein Interaction Network
nodes	proteins in the network
edges	interactions between the proteins (nodes) in the network
Cluster	Part of the PPIN that contains proteins that are more strongly connected to each other than to others. For example, proteins that are part of the same process.
RNO	Rattus Norvegicus / rat
MMU	Mus Musculus / mouse
HSA	Homo sapiens / human
HD	Huntington's Disease
HTT	Huntingtin gene
Htt	Protein coded by HTT gene
MF	GO Molecular Function
BP	GO Biological Process
FDR	False Discovery Rate
Hub	Protein in the network with a high connectivity
Enrichment Analysis	Mapping the proteins in the network to processes or functions they might be related to. The accuracy of the mapping is scored with the FDR.

2.2 Background subjects

2.2.1 Protein-Protein Interaction Networks

This thesis is all about protein-protein interaction networks, which will be shortened to PPIN. PPIN's are, as it is in the name, networks of proteins, which show the connections and interactions between proteins. These interactions can be formed by biochemical processes or electrostatic forces [Nat] and every interaction between proteins has a biological meaning. A PPIN is a mathematical display, proteins and their interactions are displayed in a graph. In a PPIN, the proteins are

represented by the nodes, the interactions and connections between proteins (nodes) in the network are represented by the edges between these nodes. The more nodes in the network, the more proteins, and the more edges between these nodes the more interactions between the proteins. Looking at PPINs can help the understanding of (complex) biological processes and systems [VFMV03]. PPINs can also help with the overall clinical picture of certain diseases. Having a better view on the proteins that are involved in the processes that cause (or are caused by) a certain disease can help the development of new therapeutics and medication for a specific disease. When, after analysis of a PPIN, it seems like a single protein(complex) plays a (massive) yet unknown role in causing the disease, a new therapy or medicine can be developed to target this specific protein(complex).

Also when comparing and analysing different PPINs, important differences and similarities can be identified. For example, looking at the same disease in different organisms can help the understanding of the disease in general. Looking at two similar diseases in humans can help the understanding of both, because it might be that the same process plays a role in both of the diseases.

Although the PPINs seem quite boring, there is a lot of underlying information behind the nodes and especially behind the edges, which can explain lots of things. As said before, in this thesis was looked at the PPINs of HD in MMU and RNO, which were compared to the PPIN of HD in humans.

2.2.2 KEGG

Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database that contains several kinds of data. The data from KEGG is composed but targeted and smaller in size than for example, STRING. KEGG contains data varying from pathways to orthologies and from diseases to enzymes. KEGG gathers data as follows: "from large-scale molecular datasets generated by genome sequencing" [Lab]. This data helps the understanding of complex functions and processes of the biological system [Lab]. A disease pathway is a sort of route that contains every protein and protein complex that is part of the disease in the right order. A pathway database contains many of these kind of pathways. In this thesis the KEGG pathway for Huntington's disease was used. The Huntington's disease pathway is available in many different organisms, but only the ones for MMU and RNO were used.

2.2.3 STRING

STRING is a database that contains data of interactions between proteins. STRING protein-protein interaction data can be noisy but there is plenty of data available. Unlike KEGG, STRING only contains proteins and their interactions, not whole disease pathways. The interactions between non-identical proteins are stored in a relational database. Every interaction is scored with the so-called STRING score. This score represents how confident STRING is that the interaction really is meaningful when knowing the current evidence [Szk21]. Whether or not a protein will be in the network depends on the chosen cut-off score. All proteins with a STRING score above the chosen cut-off score will be in the network. STRING uses a default of 0.4 as the cut-off score. Using a cut-off score is setting a certain threshold of wanted certainty for interactions present in the network.

Via Cytoscape, STRING networks can be uploaded using different kinds of queries. With the cross species query two species can be compared. The result is a network of the proteins and interactions between these two species. With the disease query, a disease can be searched for. The resulting network contains the proteins that are associated with the queried disease. The number of proteins to be visualized can be adapted manually. Using the protein query, a number of proteins can be queried. The result is a network that visualizes how these specific proteins are connected or not. At last the PubMed query can be used. With the PubMed query a query can be filled in, on which cytoscape will use textmining to match proteins and their interactions to this specific query [Con].

Another way to find STRING networks, is to go to the STRING website. On this website way more options are possible to get the wanted starting network. For example, within Cytoscape it is not possible to use the disease query on organisms other than homo sapiens. However, on the STRING website, the organisms can be changed within the disease query.

2.2.4 Cytoscape

To create the protein-protein interaction networks, the software Cytoscape was used. Cytoscape is very often used by bioinformaticians. As being said on the cytoscape website: "Cytoscape is an open source software program for visualizing molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles and other state data" [Con18]. Besides creating networks, Cytoscape can also be used to analyze networks. Many tools and apps can be downloaded from the Cytoscape AppStore to analyze the networks or to give the networks a different layout. For the network creation and analysis for this thesis, Cytoscape version 3.9.0 was used.

2.2.5 Analyzing networks

To analyse the created networks, within Cytoscape there is a tool called Analyse. This tool gives some basic topological properties of the network. These properties help to understand the structure of the network.

- Number of nodes represents the number of proteins in the network.
- Number of edges represents the number of interactions between the proteins in the network. It is the number of lines between the nodes.
- Avg number of neighbors represents the average number of connections a node has in the network.
- Network diameter represents the length of the shortest path between a node and the one that lies farthest away in the network. The network diameter says something about the connectivity of the network. When the diameter is large, the network is not that connected, compared to a small diameter, than the network is highly connected.
- Network radius represents the average number of edges it takes to travel between two proteins in the network, following the shortest path. A bigger radius means that the network is less closely connected.

- Characteristic path length is a measure of the average shortest path length between any two nodes in the network. It is a way to determine the connectivity of the network. The CPL is calculated by finding the shortest path between every pair of nodes in the network. Then taking the average of all those shortest path lengths. A network with a shorter CPL typically is more connected than a network with a larger CPL. Together with the clustering coefficient, the CPL can provide a more clear image of the structure of the network.

- Clustering coefficient represents a measure of the degree in which the nodes in the network tend to cluster together. The clustering coefficient of a node is defined as the proportion of its neighbors that are also neighbors of each other, and thus form a triangle in the network. To calculate the clustering coefficient of the whole network, the average has to be taken from the clustering coefficient of every node. A high clustering coefficient represents a network that tends to be clustered into tight groups. A low clustering coefficient represents a network that is decentralized.

- Network density represents the closeness of the network. The more interactions between the proteins the more densed the network. Calculated as the number of interactions divided by the total interactions possible in the network. The network density is a way of quantifying how crowded a network is with its connections between nodes.

- Network heterogeneity is a measure of the variation in the number of connections or the distribution of connections across the nodes in the network. One of the ways to determine the network heterogeneity is to measure the degree of each node in the network. A network has a homogeneous degree distribution if the degree of every node is similar, whereas the network has a heterogeneous degree distribution if the degrees fall within a wide range. Networks with a high heterogeneity can have more efficient communication across the network, but can be more vulnerable to failures of the most connected nodes.

- Connected components represents the number of loose groups of networks in the whole network. It is a subset of nodes that are all connected to each other by some path, but not connected to any other nodes outside of the subset.

- Analysis time represents the time it took to define these properties.

2.2.6 Clustering

Clustering is a technique which looks for groups of proteins in the network that are more connected to each other than to other (groups of) proteins in the network [WLDP10, EKGB16]. These groups are called clusters. Within PPINs there are two types of clusters. Proteins that cluster together because of interacting at the same time and same place in the cell, are called protein complexes. The other type of cluster is a functional model, in which the proteins are clustered because they are part of the same biological process, but do not interact at the same place and at the same time [WLDP10].

To cluster the networks in this thesis, the Cytoscape app "MCODE" (version 2.0.2) was used. MCODE stands for Molecular Complex Detection and is an algorithm that scans the PPIN for dense regions in the network, which may embody a protein complex. MCODE mainly is based on

the connectivity of the proteins [BH03]. When there is a group of proteins that is highly connected, it is plausible that they form a cluster.

2.2.7 Functional enrichment

When having a PPIN, and more information about a protein is wanted, for example in which processes this protein is involved in, a functional enrichment analysis can be done. Functional enrichment analysis is an analysis of the proteins to identify biological annotations that in comparison to a reference background (in this thesis the genome) are over-represented. These annotations can be used to interpret biological processes and molecular functions connected to the proteins that are being studied [GMLDVG+22]. Functional enrichment data is useful when comparing networks of different organisms. Proteins can differ among organisms, but these different proteins could be connected to the same processes or functions. Comparing functional enrichment data can lead to overlap in functions and processes which could have been missed when only looking at whether or not the same proteins are in the networks of different organisms.

When loading the functional enrichment data of the proteins in the PPIN into the network, a lot of information about the proteins is now accessible and can be used for analysis. This information contains diseases, pathways of different databases, domains, tissues, STRING clusters and three GO terms in which the proteins are involved in. The three GO terms are: Biological Process, Cellular Component and Molecular Function. Looking at one or more of these functional enrichment properties can help understanding the connections and functions of the proteins.

GO stands for Gene Ontology, and is a database that contains specific information about functions of genes [Con23]. This knowledge is used to link the proteins to the biological process, molecular function or cellular component it belongs/contributes to. When looking at a network or a cluster which contains functional enrichment data, easily can be seen which proteins are involved in the same biological process. In this thesis the focus will be on two of the three GO terms, GO Molecular Function and GO Biological Process. Every property of the functional enrichment is scored with a FDR (false discovery rate). The FDR is a score which represents the percentage of how many cases that were marked as true are actually false. This FDR is a measure of certainty, the lower the FDR the more evidence for a function of process to be connected to these proteins.

Also another method was used to find functional enrichment data of the proteins in the networks. An R-package called ClusterProfiler. According to [WHX+21], ClusterProfiler can help explore functional characteristics of gene clusters. ClusterProfiler can be used to gain knowledge of higher order functions of biological systems [YWHH12]. The data retrieved from ClusterProfiler can be easily displayed in for example dotplots. This makes it easy to see which terms are enriched in the different clusters.

3 Method

To begin this method section, the workflow followed in this thesis is shown in Figure 3. The workflow was executed twice, once for RNO *Rattus Norvegicus* and once for MMU *Mus Musculus*.

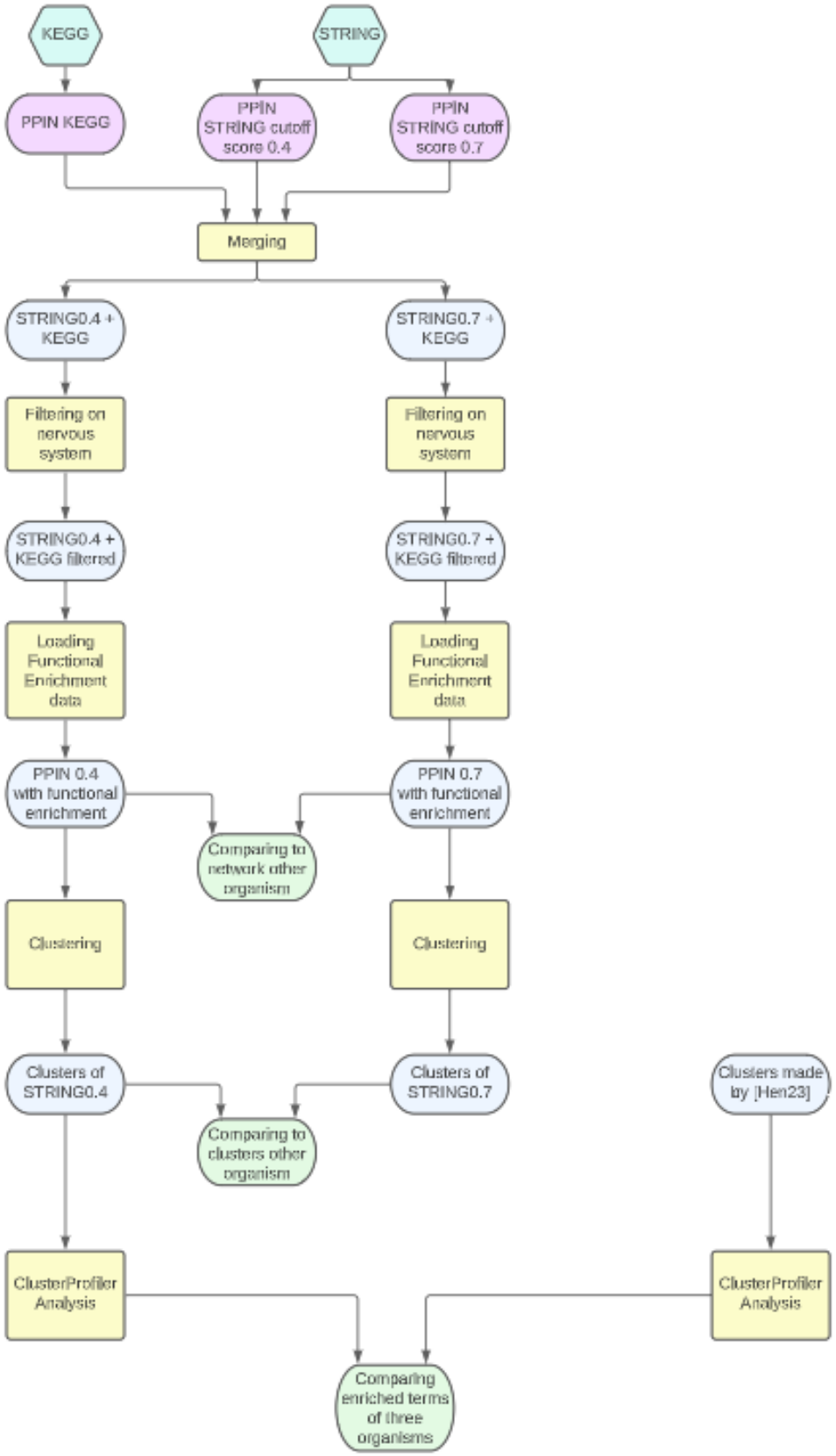

Figure 3: Workflow followed. Purple means an unmodified network, blue represents a created network, green means comparing between two or more networks and yellow means a process in the method.

3.1 Gathering data

In order to create the networks, first the data was gathered. For the networks to be as complete as possible, different kinds of data were used. To start the creation of a network, a base network was needed. The base networks in this thesis were STRING networks. For many organisms there is a network available on the STRING website for the proteins associated with HD. On the STRING website the disease query was used to find the network for proteins associated to HD together with the wanted organism. The created networks were exported to Cytoscape.

A disease pathway represents how and which proteins or protein-complexes are connected to the specific disease. In this thesis the KEGG disease pathways of HD for MMU and RNO were used. The .KGML file containing the pathway was uploaded into Cytoscape. However, the KEGG network needed to be modified in order to be able to merge the network to the STRING network. The names of the proteins were converted to UniprotIds and the protein complexes were split into single proteins, however, retaining every interaction between these proteins. Figure 13 and Figure 22 show the KEGG pathways of MMU and RNO respectively. To expand the pathways, the script written by Aster de Boer [dB23] was used. This script splits every complex in single proteins, reconnecting the connected proteins in this complex after the split. In addition to this splitting the script also adds the Uniprot Ids to the network. This script resulted in a .gml file that contains the network and was uploaded to Cytoscape. Because of [dB23] using the human pathway for Parkinson (hsa05010), wherever there is hsa05010 in the script, this was changed to ...05016, with on the three dots the abbreviation for the wanted organism. In this thesis these were mmu and rno.

3.2 Network creation

Per organism two networks were created, one with a cut-off score of 0.4 (Figure 11 for MMU and Figure 20 for RNO) and one with a cut-off score of 0.7 (Figure 12 for MMU and Figure 21 for RNO) using STRING. The cut-off score is used as a certainty threshold for the STRING score of the interaction between two proteins. When using the same query with a higher cutoff score, the interactions occurring between the proteins in the network are more certain to happen, which can lead to new insights. A cut-off score of 0.4 means that roughly four out of ten interactions in the network truly happen. A cut-off score of 0.7 means that roughly seven out of ten interactions in the network truly happen. So there are possible less proteins and interactions in the 0.7 cut-off score network, however, these interactions are more likely to happen than when choosing a lower cut-off score.

These STRING networks were exported to Cytoscape to be able to merge them to other networks and to analyze them later in the process. The original KEGG pathway can be loaded into Cytoscape to see which proteins form complexes and to look at the structure of the network. As mentioned before, to be able to merge the pathway to the STRING network, the pathway needs to be expanded. The original pathways are visible in Figure 13 and Figure22 for MMU and RNO respectively.

The expanded networks were merged to the STRING networks, using the Merge tool within Cytoscape. The column of the pathway table containing the uniprot ids, was set as the key-column. Using the merge tool in Cytoscape, there was chosen for Union as merging type, and the columns

containing the Uniprot Ids were chosen as attributes that should match. The merged networks for MMU are in Figure 14 and Figure 15 for 0.4 and 0.7 as cut-off score respectively. For RNO the merged networks are visible in Figure 23 and Figure 24 for 0.4 and 0.7 as cut-off score respectively.

The merged networks now were filtered on whether or not there is a connection to the nervous system. Because symptoms of HD mainly effects brain activity, looking at proteins associated to the nervous system is the most effective. If there is no value in the column 'Nervous System' there is no connection to it, so filtering on a value in this column gave a better view on proteins associated with brain activities. The merged networks that were filtered on nervous system for MMU are in Figure 16 and Figure 17 for 0.4 and 0.7 as the cut-off score respectively. For RNO the filtered on nervous system networks are visible in Figure 25 and Figure 26 for 0.4 and 0.7 as the cut-off score respectively.

3.3 Functional enrichment

In order to analyze the functions of the proteins in the networks, functional enrichment data was loaded into the networks within Cytoscape. In this thesis, STRING Enrichment App was used and the functional enrichment was loaded using the genome as the background network and was focused on the GO Molecular Function and GO Biological Process. The enriched terms obtained by STRING Enrichment app were scored with an FDR. The FDR is a score which represents the percentage of how many cases that were marked as true are actually false. This FDR is a measure of certainty, the lower the FDR the more evidence for a function of process to be connected to these proteins. The lower the FDR the more certain the process or function is related to the proteins in the network. So the processes and functions with the lowest FDR are on top of the list.

Together with the function within cytoscape, also an R-package was used to search for the functional enrichment. This package is called clusterProfiler and is a part of BiocManager. Below, the R-script used.

```
if (!require("BiocManager", quietly= TRUE))
  install.packages("BiocManager")
install.packages("readxl")
BiocManager::install("clusterProfiler")
BiocManager::install("org.Mm.eg.db")
BiocManager::install("org.Rn.eg.db")
BiocManager::install("org.Hs.eg.db")
library("clusterProfiler")
library("org.Mm.eg.db")
library("org.Rn.eg.db")
library("org.Hs.eg.db")
library("readxl")

compare <- function(data, org){
  datanew <- read_excel(data)
  ck <- compareCluster(geneCluster=datanew, fun="enrichGO", OrgDb=org,
   pvalueCutoff=0.01)
  return(dotplot(ck, showCategory=3))}

rnoclus4 <- compare(file.choose(), "org.Rn.eg.db")
```

```
mmuclus4 <- compare(file.choose(), "org.Mm.eg.db")
hsaclus <- compare(file.choose(), "org.Hs.eg.db")
```

The data is in this case an excel file with a column per cluster or network containing the GeneIds of the genes in the cluster or network. Because of the fact that in the networks only uniprot ids are available, these uniprot ids had to be converted to GeneId, using the Uniprot ID mapping tool on the following website. For the different organisms, different databases must be used. The so-called OrgDbs. These are available in Bioconductor and for many organisms. These OrgDbs must be loaded as a package, but the advantage is that these packages are updated biannualy [WHX+21]. The function returns a dotplot with the top-3 enrichment results. As seen in the script, in this thesis the compareCluster function was run multiple times to compare the enrichment of the different organisms. This had to be done separately because only one organism database can be chosen per run.

3.4 Clustering

In order to identify possible protein complexes in the networks, the networs were clustered using MCODE clustering. As described before, the MCODE clustering algorithm searches for groups of proteins that are more closely connected to each other than to other proteins [BH03]. Within Cytoscape the MCODE application with version 2.0.2 was used. The default parameters were used, which are shown in Table 2. The obtained clusters were exported as new (sub)networks in the collection. The clusters were then analyzed using the above R-script. This resulted in a dotplot per organism which visualizes the top-3 enriched terms per cluster.

Table 2: Parameters used for MCODE clustering.

Parameter	Value
Find clusters	In whole network
Include Loops	No
Degree cutoff	2
Haircut	Yes
Fluff	No
Node score cutoff	0.2
K-Core	2
Max. Depth	100

4 Results

4.1 Creating networks

The result was two networks per organism, which are the merged networks that were filtered on whether or not there is a connection to the nervous system. One with a cut-off score of 0.4 (Figure 16 for MMU and Figure 25 for RNO) and another one with a cut-off score of 0.7 (Figure 17 for

MMU and Figure 26 for RNO). During the process of creating these networks, the analyze tool within Cytoscape was used three times. The pure STRING networks, the pathway network merged with the STRING network, and the filtered on Nervous system network were analyzed. These results were per organism and per cut-off score displayed in the following tables: MMU with cut-off score 0.4 in Table 3; MMU with cut-off score 0.7 in Table 4; RNO with cut-off score 0.4 in Table 5 and RNO with cut-off score 0.7 in Table 6.

Within these tables, some differences in the networks were visible. Merging KEGG with the STRING network for all networks increased the number of nodes and also the number of edges, but it decreased the average number of neighbors. However, filtering the merged network on whether or not having a score for nervous system decreased both the number of nodes and number of edges, but increased the average number of neighbors. The increase of nodes and edges in the networks after the merge was logical. The KEGG pathway surely contained proteins and interactions that were not in the STRING network, thus increased the number of nodes and edges. When adding new proteins and interactions to an existing network there is a major chance of these proteins having less neighbors than the average in the existing network. This decreased the average number of neighbors of the merged network. When filtering on whether or not there was a interaction to the nervous system, some proteins were deleted from the network, and thereby also the edges connected to this node. It is possible that the proteins connected to the nervous system are more closely connected to each other than proteins with out this connection. Deleting these non-connecting proteins with possible a low degree may increase the average number of neighbors of the new network.

When looking at the number of connected components in the networks, there were some things that stood out when looking at both MMU tables and the 0.4 RNO table. The number of connected components increased massively when looking at the STRING network and the merged one. This was caused by the pathway network. In the pathway network there were also compounds which were not connected to the whole network. These non-connected compounds and some paths are part of the network but are no proteins. Since the goal was to create protein-protein interaction networks, these were filtered out. These non-connected components disappeared when the networks were filtered on nervous system. When looking at the STRING RNO 0.7 network in Figure 21 and the merged one in Figure 15, there were in the top right corner two non-connected components. After filtering these were gone visible in Figure 17, but in the network were still 2 non-connected components, which were not that visible, but still present. These proteins were connected to the nervous system but had no interactions to the other proteins in the network.

The trend in the diameter and radius of the networks was consistent between the networks. The values did not differ much in the three networks. The rest of the topological properties did not differ that much between the three networks. However, there was still some fluctuation between the values. The values in the STRING and filtered column were pretty equal most of the time, with a small increase/decrease of the value in the merged column.

When looking at the same organisms but a different cut-off score some things stood out. The number of nodes was equal, but the number of edges in the network was less in the 0.7 network, even as the average number of neighbors. The characteristic path length, the network density and the network centralization were lower in the 0.7 network, however the network heterogeneity and

clustering coefficient of the 0.7 network were higher than the 0.4 network. These differences can be explained: in the 0.7 network were less interactions between the proteins because of the higher needed string score, this made the network less dense, but more heterogeneous (more nodes with the same or similar degree). The density of the network is coupled to the centralization, so a lower density also means less centralization. When there are less interactions, the characteristic path length increases.

Table 3: Table shows results of MMU network with cut-off score 0.4 using the Analyze tool on different moments in the process. The "Figure .." represents which network was analyzed in the column.

MMU 0.4 cut-off	STRING (Figure 11)	STRING + KEGG (Figure 14)	Filtered on Nervous system (Figure 16)
Number of nodes	296	351	295
Number of edges	6582	8870	8483
Avg number of neighbors	44.473	42.120	45.098
Network diameter	4	5	4
Network radius	2	3	2
Characteristic path length	2.213	2.379	2.206
Clustering coefficient	0.735	0.719	0.747
Network density	0.151	0.126	0.153
Network heterogeneity	0.700	0.750	0.681
Network centralization	0.254	0.232	0.253
Connected components	1	18	1
Analysis time (sec)	0.137	0.118	0.137

Table 4: Table shows results of MMU network with cut-off score 0.7 using the Analyze tool on different moments in the process. The "Figure .." represents which network was analyzed in the column.

MMU 0.7 cut-off	STRING (Figure 12)	STRING + KEGG (Figure 15)	Filtered on Nervous system (Figure 17)
Number of nodes	296	351	295
Number of edges	4751	7039	6658
Avg number of neighbors	32.101	31.617	33.214
Network diameter	6	6	5
Network radius	3	3	3
Characteristic path length	2.850	2.961	2.819
Clustering coefficient	0.735	0.740	0.775
Network density	0.109	0.095	0.113
Network heterogeneity	0.816	0.822	0.774
Network centralization	0.208	0.195	0.208
Connected components	1	18	1
Analysis time (sec)	0.073	0.091	0.070

Table 5: Table shows results of RNO network with cut-off score 0.4 using the Analyze tool on different moments in the process. The "Figure .." represents which network was analyzed in the column.

RNO 0.4 cut-off	STRING (Figure 20)	STRING + KEGG (Figure 23)	Filtered on Nervous system (Figure 25)
Number of nodes	283	382	280
Number of edges	6488	8820	7899
Avg number of neighbors	45.852	40.717	45.607
Network diameter	4	6	4
Network radius	2	3	2
Characteristic path length	2.190	2.528	2.188
Clustering coefficient	0.705	0.686	0.714
Network density	0.163	0.114	0.163
Network heterogeneity	0.694	0.812	0.687
Network centralization	0.218	0.196	0.218
Connected components	1	25	1
Analysis time (sec)	0.820	0.257	0.169

Table 6: Table shows results of RNO network with cut-off score 0.7 using the Analyze tool on different moments in the process. The "Figure .." represents which network was analyzed in the column.

RNO 0.7 cut-off	STRING (Figure 21)	STRING + KEGG (Figure 24)	Filtered on Nervous system (Figure 26)
Number of nodes	283	329	280
Number of edges	4678	6885	6128
Avg number of neighbors	33.295	32.881	33.561
Network diameter	6	6	6
Network radius	4	4	4
Characteristic path length	2.946	2.812	2.938
Clustering coefficient	0.747	0.739	0.769
Network density	0.119	0.101	0.121
Network heterogeneity	0.809	0.848	0.787
Network centralization	0.218	0.216	0.216
Connected components	3	3	3
Analysis time (sec)	0.060	0.148	0.077

4.2 Functional enrichment

4.2.1 Cytoscape

In order to have a better overview of the enriched terms per category, the table was filtered on GO Biological Process and GO Molecular Function separately. When the wanted category was selected the number the number of rows in the table was made visible, which represents the number of different, in this case, GO Molecular Functions and GO Biological Processes. The networks for MMU contained 111 different GO Molecular Functions and 334 GO Biological Processes. The

networks for RNO contained 98 different GO Molecular Functions and 352 GO Biological Processes.

Within Cytoscape, a color can be assigned to the enriched terms of interest in order to see which proteins in the network are connected to these enriched terms. In Table 7 and Table 8 the six enriched terms with the lowest FDR were given of the GO Molecular Function and the GO Biological Process respectively. Not the best 5 but the best 6 terms were given in the tables because of the fact that whether looked at the MF or the BP, the first six were the same in all of the four networks, however, not in the same order. In Table 7 can be seen that the order of the MFs was the same for the two networks of MMU, and the same for the two networks of RNO. However, when comparing the organisms, the same terms were enriched but in a different order. The same applies to Table 8. The two networks of MMU had the same BPs in the same order and the two networks of RNO also had the same BPs in the same order. Again when comparing both organisms, the biological processes were the same in all of the four networks, but had different orders for RNO and MMU.

Keeping in mind the hierarchy of GO, the difference in order did matter. The hierarchy of the GO Molecular Functions is as follows: the two NADH dehydrogenase activities are electron transfer activities, which again is a oxidoreductase activity which again is a catalytic activity. The fact that for RNO the NADH dehydrogenase activities were above catalytic activity in the list made that the proteins were more related to these two specific catalytic activities than to catalytic activities in general. For MMU this was reversed. General catalytic activity was more important than the two specific catalytic activities.

The hierarchy of the GO Biological Processes is as follows: oxidative phosphorylation is a cellular respiration which together with electron transport chain is a generation of precursor metabolites and energy. The fact that for both of the organisms oxidative phosphorylation was the process to which the proteins were mostly related to, makes that it was higher on the list than the generation of precursor metabolites and energy in general. The proteins in the network of RNO were more closely related to electron transport chain than to the generation of precursor metabolites and energy in general. When looking at these two processes in the MMU networks these two were reversed. So the proteins in the network of MMU were more related to the generation of precursor metabolites and energy in general. In the two sections below the six GO MFs and GO BPs were described.

4.2.2 GO Molecular Functions lowest FDR (Cytoscape)

In this section some extra information was provided about the six GO Molecular Functions with the lowest FDR which are in Table 7: Catalytic activity; NADH dehydrogenase activity; NADH dehydrogenase (ubiquinone) activity; Electron transfer activity; Oxidoreductase activity and Proton transmembrane transporter activity. Keep in mind the hierarchy described above in Section 4.2.1.

- Catalytic activity is also known as enzyme activity. Works as a catalysator of biochemical reactions, to be able to let the enzyme bind to the substrate [EE09].
- NADH dehydrogenase activity makes sure the reaction to form NAD^+ from NADH is catalyzed [EE02].

Table 7: Showing the ids and GO term of the six GO Molecular Functions with the lowest FDR of the 4 networks. The enriched terms are the same in all of the four columns, however when looking at the different organisms they have a different order.

	Merged MMU 0.4	Merged MMU 0.7	Merged RNO 0.4	Merged RNO 0.7
1st MF	Catalytic activity (GO:0003824)	Catalytic activity (GO:0003824)	NADH dehydrogenase activity (GO:0003954)	NADH dehydrogenase activity (GO:0003954)
2nd MF	NADH dehydrogenase activity (GO:0003954)	NADH dehydrogenase activity (GO:0003954)	NADH dehydrogenase (ubiquinone) activity (GO:0008137)	NADH dehydrogenase (ubiquinone) activity (GO:0008137)
3rd MF	NADH dehydrogenase (ubiquinone) activity (GO:0008137)	NADH dehydrogenase (ubiquinone) activity (GO:0008137)	Catalytic activity (GO:0003824)	Catalytic activity (GO:0003824)
4th MF	Electron transfer activity (GO:0009055)	Electron transfer activity (GO:0009055)	Oxidoreductase activity (GO:0016491)	Oxidoreductase activity (GO:0016491)
5th MF	Oxidoreductase activity (GO:0016491)	Oxidoreductase activity (GO:0016491)	Electron transfer activity (GO:0009055)	Electron transfer activity (GO:0009055)
6th MF	Proton transmembrane transporter activity (GO:0015078)	Proton transmembrane transporter activity (GO:0015078)	Proton transmembrane transporter activity (GO:0015078)	Proton transmembrane transporter activity (GO:0015078)

- NADH dehydrogenase (ubiquinone) activity catalyzes the following reaction: NADH + ubiquinone + 5 $H^+ \Leftrightarrow NAD^+$ + ubiquinol + 4 H^+ [EE22].
- Electron transfer activity makes sure that an electron donor can donate one or more electrons to one or more electron acceptors. The term electron transfer activity should only be used when there is a transmembrane electrochemical gradient generated during the transfer [EE17].
- Oxidoreductase activity catalyzes redox reactions [EE08c].
- Proton transmembrane transporter activity makes sure that protons can transfer over the membrane from one site to the other [EE18].

4.2.3 GO Biological Processes lowest FDR (Cytoscape)

In this section some extra information was provided about the six GO Biological Processes with the lowest FDR which are in Table 8: Oxidative phosphorylation; ATP metabolic process; Generation of precursor metabolites and energy; Electron transport chain; Respiratory electron transport chain and Cellular respiration. Keep in mind the hierarchy described in Section 4.2.1.

Table 8: Showing the ids and GO term of the six GO Biological Processes with the lowest FDR of the 4 networks. The enriched terms are the same in the four columns, however in a different order when comparing the two organisms.

	Merged MMU 0.4	Merged MMU 0.7	Merged RNO 0.4	Merged RNO 0.7
1st BP	Oxidative phosphorylation (GO:0006119)	Oxidative phosphorylation (GO:0006119)	Oxidative phosphorylation (GO:0006119)	Oxidative phosphorylation (GO:0006119)
2nd BP	ATP metabolic process (GO:0046034)	ATP metabolic process (GO:0046034)	Electron transport chain (GO:0022900)	Electron transport chain (GO:0022900)
3rd BP	Generation of precursor metabolites and energy (GO:0006091)	Generation of precursor metabolites and energy (GO:0006091)	Generation of precursor metabolites and energy (GO:0006091)	Generation of precursor metabolites and energy (GO:0006091)
4th BP	Electron transport chain (GO:0022900)	Electron transport chain (GO:0022900)	Respiratory electron transport chain (GO:0022904)	Respiratory electron transport chain (GO:0022904)
5th BP	Respiratory electron transport chain (GO:0022904)	Respiratory electron transport chain (GO:0022904)	Cellular respiration (GO:0045333)	Cellular respiration (GO:0045333)
6th BP	Cellular respiration (GO:0045333)	Cellular respiration (GO:0045333)	ATP metabolic process (GO:0046034)	ATP metabolic process (GO:0046034)

- Oxidative phosphorylation is the process in which ADP is phosphorylated to ATP which is accompanied by the respiratory chain that oxidates a metabolite. A gradient over the membrane is established by oxidating a compound [EE08b].
- ATP metabolic process represents all the chemical reactions and pathways involved in the metabolism of ATP [EE08d].
- Generation of precursor metabolites and energy is the process that makes and extracts energy from energy containing substances called precursor metabolites [EE08a].
- Electron transport chain transfers electrons from electron donors to electron acceptors [EE21].
- Respiratory electron transport chain is the process in which electrons are tranferred from electron donors as NADH and $FADH_2$ to several donor acceptors, which will establish a transmembrane electrochemical gradient [EE10].
- Cellular respiration is the release of energy extracted from organic and inorganic resources, which can take place with or without oxygen [EE14].

4.2.4 ClusterProfiler

ClusterProfiler was used to compare the whole networks with the different cut-off scores per organism. To see if there was any difference in the enrichment of the two model organisms. Figure 4 shows the results of the MMU networks, Figure 5 shows the results of the RNO networks. Both of

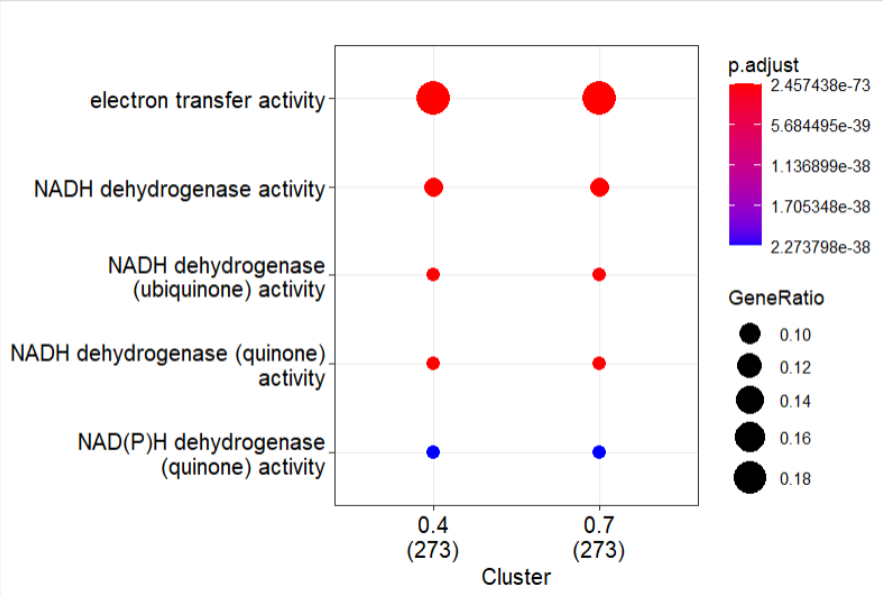

Figure 4: CompareCluster applied on the unclustered networks of MMU with cut-off score 0.4 and 0.7. Figure shows the five enriched terms of both of the networks for MMU.

the organisms had the same amount of input ids in the different cut-off scores. And in addition to this, both of the organisms had the same 5 enriched terms in both of the networks. When looking at the enriched terms in the different organisms could be seen that these were the same 5 terms. Also could be seen that these results differ from the results obtained by Cytoscape.

4.2.5 Enriched terms connected to HD

According to [JB12], causes the mutation in the huntingtin gene neuronal dysfunction which will lead to cell death. The neuronal dysfunction is among other things caused by mitochondrial dysfunction. The mutant HTT effects the mitochondria negatively in at least three different ways. First it damages a mitochondrion which will now produce reactive oxygen species (ROS). The fact that now the damaged mitochondrion produces ROS, will only damage more and more mitochondria. An imbalance in these ROS will cause oxidative stress, which in turn causes the ROS to damage among other things DNA and proteins. The damage caused by these ROS has been linked to neurodegeneration [RHT, SZK$^+$15]. Second, it causes mitochondria to split more often and makes them less movable. As a third thing, the mutant HTT causes a reduction of transcription and expression of PGC-1a, which leads to down-regulation of genes that are involved in mitochondrial bio-genesis [JB12]. Also down-regulation of PGC-1a causes an impaired response of the mitochondria to antioxidants [SZK$^+$15]. PGC-1a has the same function in mice [CAS$^+$] and in rats [AZX$^+$22].

All of the above GO Biological Processes and GO Molecular Functions are connected to the mitochondria or take place within the mitochondria. Tissues that consume a lot of energy, for example the central nervous system or muscles, usually contain a large concentration of mitochondria. Mitochondria have among other things the function to regulate the apoptosis of a cell [OS01]. The neurons in the central nervous system are quite fragile. They are not capable to store glycogen

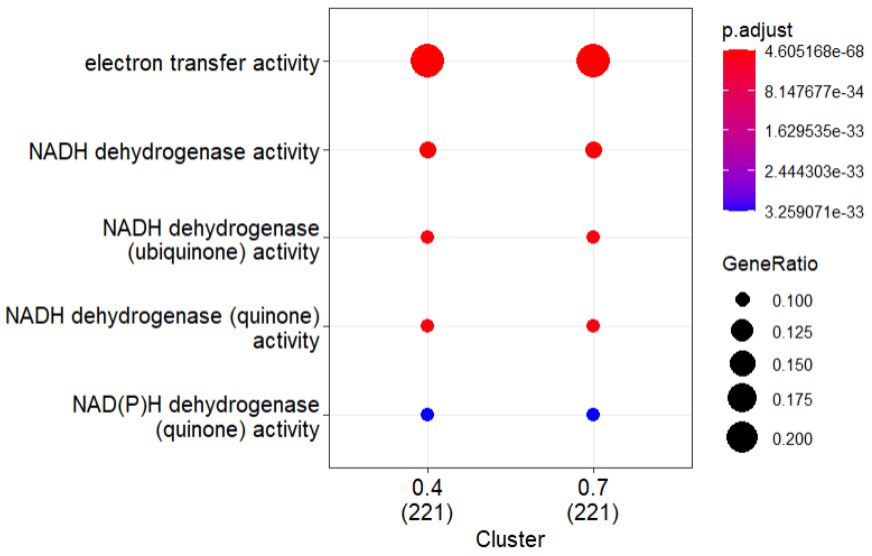

Figure 5: CompareCluster applied on the unclustered networks of RNO with cut-off score 0.4 and 0.7. Figure shows the five enriched terms of both of the networks for RNO.

and use one-fifth of the total amount of oxygen available. For a neuron to function well it needs a sufficient amount of oxygen and glucose available at every moment. According to [ANK+13] mitochondrial involvement in HD is suggested when there are patients with higher concentrations of lactate and a reduced metabolism of glucose in many areas in the brain. Besides that, a neuron is very sensitive and can handle only a small amount of antioxidants, and only a low level of mitochondrial stress [SZK+15]. Exceeding these levels can lead to neuronal dysfunction. In addition to this, damage to mitochondria in muscle cells can lead to involuntary movements which are one of the symptoms of HD.

4.3 Clustering

Using MCODE on the 4 filtered networks resulted in a number of clusters. For MMU, the number of clusters of the networks filtered on nervous system with cut-off score 0.4 and 0.7 were 10 and 21 respectively, ranging form 86 to 3 proteins in a cluster. For RNO, the number of clusters of the networks filtered on nervous system with cut-off score 0.4 and 0.7 were 8 and 17 respectively, ranging from 86 to 3 proteins in a cluster. The fact that the 0.7 networks had more clusters can be declared by the increase of the clustering coefficient in the networks, as can be seen in Table 4 and Table 6. Because of the decrease in interactions, the clustering coefficient increases which may lead to more clusters in a network with a higher cut-off score. Because of the functional enrichment data loaded into the PPINs before these were clustered, the information was also available in the found clusters. Clusters often represent proteins that are involved in the same process or have the same function, however, there is some bias towards proteins that physically interact in a protein complex. Using this functional enrichment data, the function of certain clusters can be easily identified [WLDP10].

The clusters of interest were the ones which contained a protein which was a first neighbor of the Htt protein in the networks with a cut-off score of 0.4. This resulted in the following clusters of interest for MMU: Cluster 1 to Cluster 9 (so nine clusters) and resulted in the following clusters for RNO: Cluster 1 to Cluster 7 (so seven clusters). The clusters of interest were analyzed using clusterProfiler to look at which terms were enriched in the different clusters. The results of these analyses are visible in Figure 6 for MMU and in Figure 7 for RNO.

When looking at Figure 6 can be seen that most clusters had three unique enriched terms. However there were three clusters that were connected to motor activity and microtubule motor activity. These were cluster 3, 4 and 9 for MMU. In Figure 7 can be seen that cluster 3 and 4 both had motor activity as enriched term and that cluster 2 and 6 both had endopeptidase activity as enriched term.

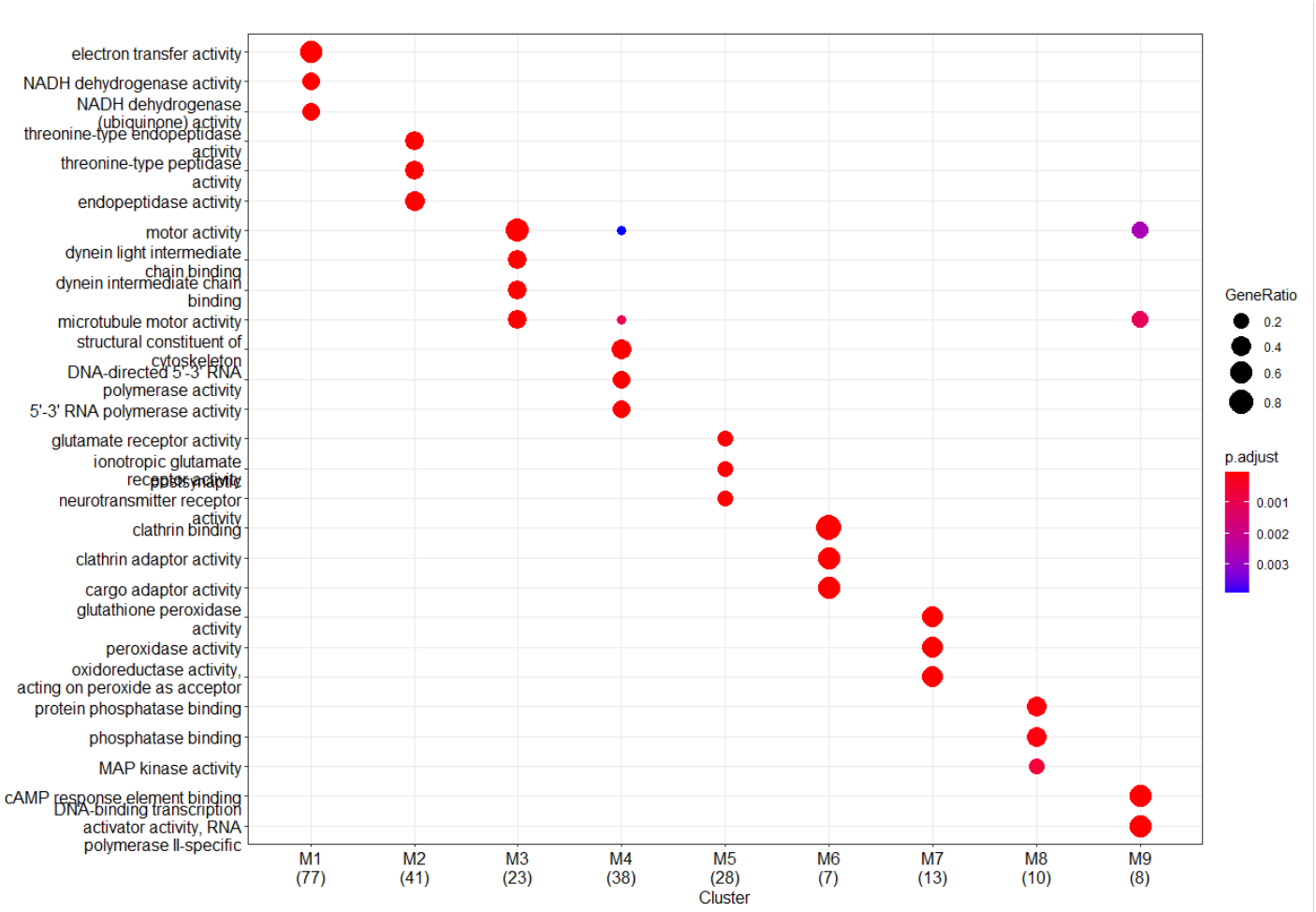

Figure 6: Per cluster of interest for MMU the top-3 enriched terms are showed. Some functions are connected to multiple clusters. Cluster 3, 4 and 9 are connected to motor activity and microtubule motor activity.

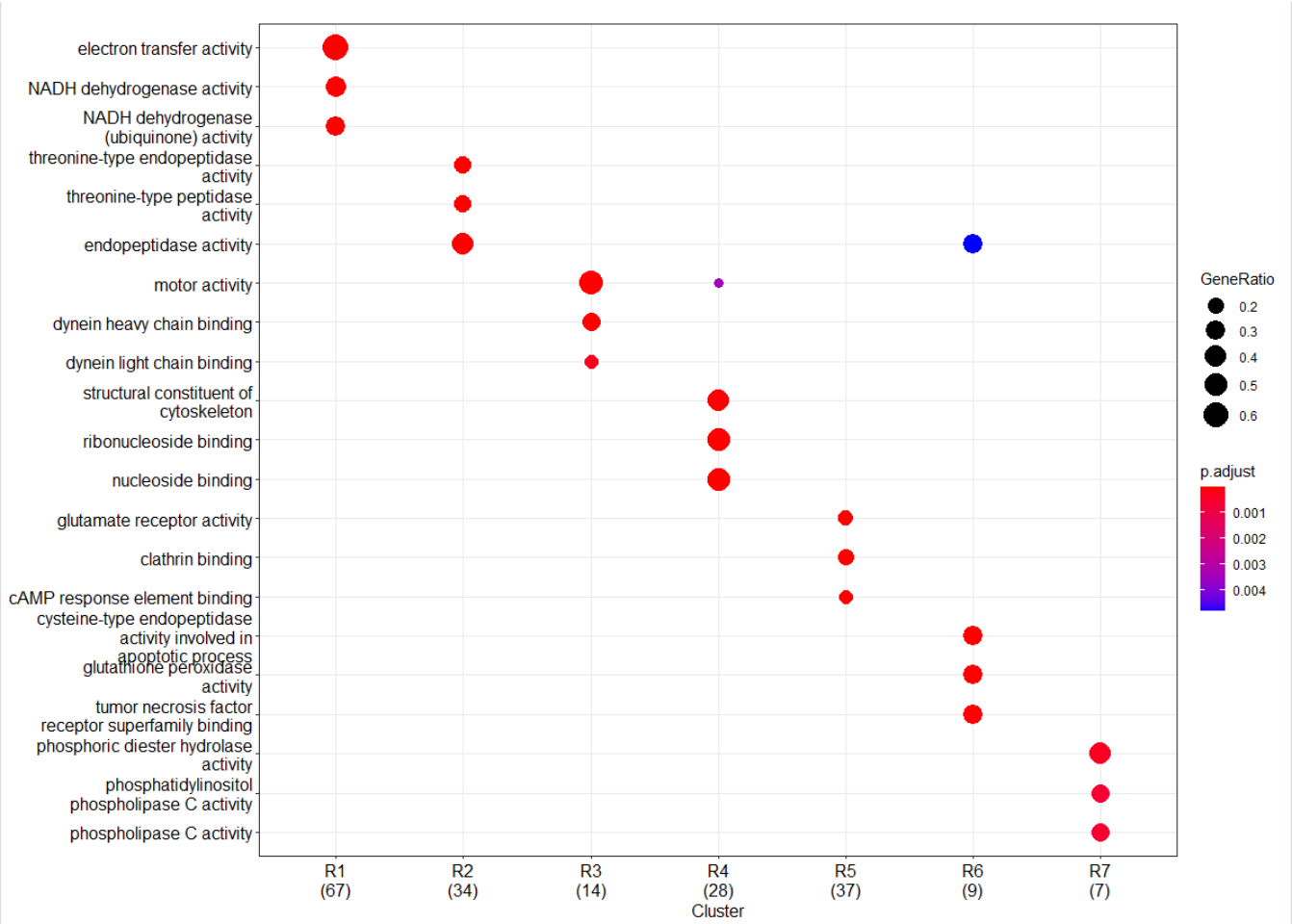

Figure 7: Per cluster of interest for RNO the top-3 enriched terms are showed. Some functions are connected to multiple clusters. So is motor activity connected to both cluster 3 and 4, and is endopeptidase activity connected to both cluster 2 and 6.

5 Comparison

The results in Section 4 showed a high similarity between the clusters obtained using a cut-off score of 0.4 and 0.7 for both of the model organisms. Looking at the human data [Hen23], only data was available with a cut-off score of 0.4. So with the comparison of the three organisms, for RNO and MMU only the clusters obtained using the 0.4 cut-off data were used. In the subsections below, some comparisons were done between the different organisms.

5.1 RNO and MMU

In order to compare the networks of RNO and MMU created in this thesis, was looked at both cut-off scores to be complete. The comparison was split into two sections, a section in which the clusters were compared, and a second section in which the functional enrichment of the different organisms were compared. In all of the networks for RNO and MMU, Htt was in the middle of the network, however, this was not the protein with the highest degree.

5.1.1 Clusters

The clusters of RNO and MMU were compared per cut-off score and the comparisons are visible in Figure 8 for 0.4 and in Figure 9 for 0.7. There was looked at the number of proteins in the clusters of both organisms and how many of these proteins overlapped. The blue bars represent the number of proteins in the clusters of RNO, the orange bars represent the number of proteins in the clusters of MMU and the grey bar is the number of proteins that is present in the clusters of both MMU and RNO. The numbers below the bars are the clusters that were compared. The first number is the number of the RNO cluster, the second number is the number of the MMU cluster.
In both of the comparisons, the second group of bars was the most remarkable. The clusters only differed one protein in count, but the number of proteins that overlapped is one less than the number of proteins in the smallest cluster. Cluster 2 from the 0.4 cut-off score networks in both organisms were the only clusters that contained Htt. Cluster 2 in RNO, had 43 nodes and 955 edges. Cluster 2 in MMU had 42 nodes and 1011 edges. The huge overlap between the proteins in the second cluster of the two model organisms together with the fact that these clusters were the only two clusters containing Htt is remarkable. Both clusters contained mostly proteins that are subunits of the Proteasome 20S. In addition to this both clusters contained TP53. The proteasome is part of the ATP-dependent proteolytic pathway and degrades most proteins. The 20S proteasome uses a threonine active site for peptidase activities to function [CTG96].

Both clusters had the same three enriched terms according to the clusterProfiler analysis: threonine-type peptidase activity; threonine-type endopeptidase activity and endopeptidase activity. Also both clusters had the same enriched terms looking at the GO Molecular Functions and GO Biological Processes obtained by Cytoscape in Table 9. How the proteins were connected to the GO Molecular Functions is visible in Figure 18 and Figure 27 for MMU and RNO respectively. How the proteins were connected to the GO Biological Processes is visible in Figure 19 and Figure 28 for MMU and RNO respectively. These enriched terms obtained by Cytoscape were partly overlapping with the results obtained by clusterProfiler. Threonine-type endopeptidase activity and endopeptidase activity came up in both analyzes. In addition to this, the result obtained by Cytoscape showed proteasomal activities, which is logical because almost every protein in the cluster is a subunit of the 20S Proteasome.

Table 9: Ids and GO terms of the GO Molecular Functions and GO Biological Processes according to Cytoscape of both second clusters of MMU and RNO.

GO Molecular Function	GO Biological Process
Threonine-type endopeptidase activity (GO:0004298)	Protein catabolic process (GO:0030163)
Endopeptidase activity (GO:0004175)	Ubiquitin-dependent protein catabolic process (GO:0006511)
Hydrolase activity (GO:0016787)	Proteasome-mediated ubiquitin-dependent protein catabolic process (GO:0043161
Proteasome-activating ATPase activity (GO:0036402)	Proteasomal ubiquitin-independent protein catabolic process (GO:0010499)
Porin activity (GO:0015288)	Catabolic Process (GO:0009056)

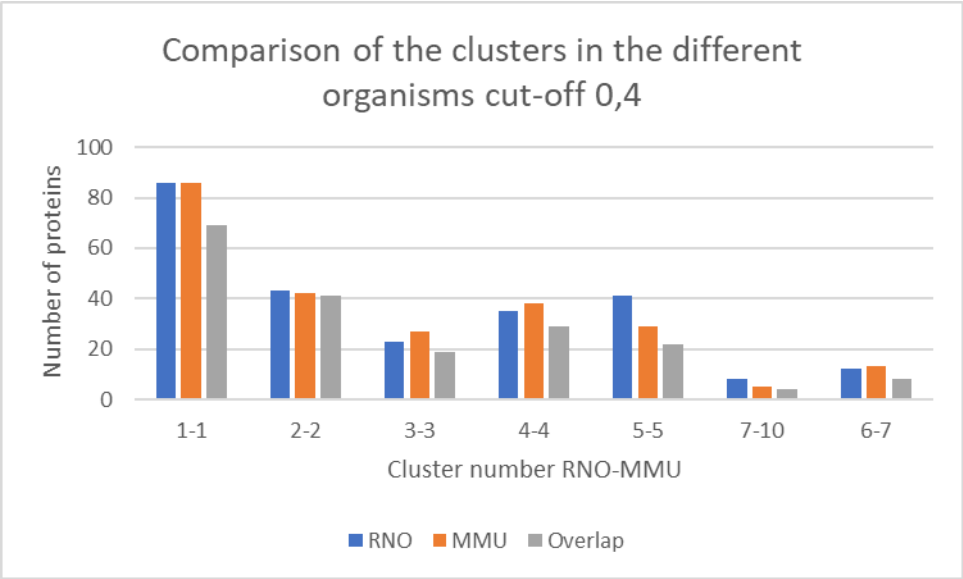

Figure 8: Comparison of the clusters between the different organisms. Cut-off score of 0.4. The number below the bars is the number of the clusters. So number of RNO cluster - number of MMU cluster. The blue and orange bars represent the number of proteins that are in the clusters and the grey is the number of the proteins that overlap between those two clusters.

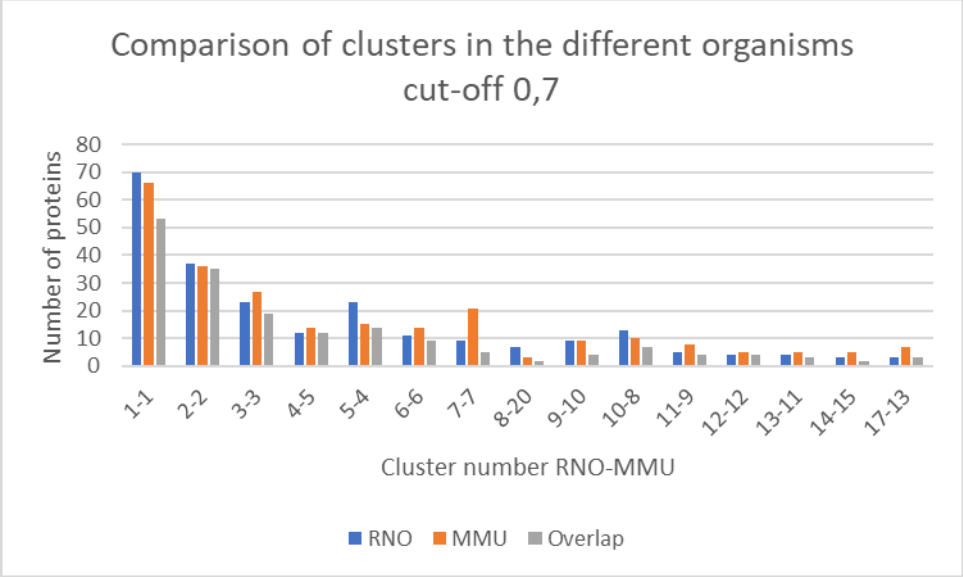

Figure 9: Comparison of the clusters between the different organisms. Cut-off score of 0.7. The number below the bars is the number of the clusters. So number of RNO cluster - number of MMU cluster. The blue and orange bars represent the number of proteins that are in the clusters and the grey is the number of the proteins that overlap between those two clusters.

5.1.2 Functional enrichment

In both of the results of the functional enrichment analysis, using Cytoscape and ClusterProfiler there was only a small difference between MMU and RNO when looking at the whole networks. When looking at the Cytoscape analysis, the same 6 terms were enriched for both MF (Table 7) and BP (Table 8), however, in a different order for the organisms. But when looking at the results of the clusterProfiler analysis in Figure 4 for MMU and Figure 5 for RNO, the same 5 terms were enriched in the same order too. Using clusterProfiler the networks also were analyzed per cluster, to see the enriched terms per cluster. The first seven terms of both lists were the same in the same order. So the first two clusters had the same enriched terms in both organisms. When looking at the rest of the lists in total eleven of the terms occurred in both organisms.

- GO:0009055 – electron transfer activity
- GO:0003954 – NADH dehydrogenase activity
- GO:0008137 – NADH dehydrogenase (ubiquinone) activity
- GO:0070003 – threonine-type peptidase activity
- GO:0004298 – threonine-type endopeptidase activity
- GO:0004175 – endopeptidase activity
- GO:0003774 – motor activity
- GO:0008066 – glutamate receptor activity
- GO:0030276 – clathrin binding
- GO:0035497 – cAMP response element binding
- GO:0004602 – glutathione peroxidase activity

When looking at the above enriched terms, they can be linked to HD. NADH dehydrogenase activity, NADH dehydrogenase (ubiquinone) activity and electron transport chain are part of the energy metabolism within mitochondria. The mutant HTT disrupts the energy metabolism in the mitochondria, which has a negative effect on, among others, these parts of the metabolism [GGM$^+$96]. Also motor activity can be coupled to HD. Shortfall of motor activity can be declared by striatal damage caused by mutant HTT [DB18]. One of the symptoms of HD is involuntary movement.

The mutant HTT with thus more CAG repeats disrupts the CRE transcription. The CRE pathway plays an important role in the nervous system for cells to survive [SR03]. The mutant HTT makes sure that CREB cannot bind to the cAMP response elements, which ensures a decreased transcription of CRE [SR03]. Glutathione peroxidase is an antioxidant enzyme that maintains the amount of ROS in cells [GPY$^+$85]. According to [PSSPC$^+$04] there was in their model organisms an increase in amount of ROS, however no change in the activity of glutathione peroxidase. So mutant HTT causes an increase in ROS in a cell however glutathione peroxidase is not capable

of bringing this amount back to a healthy level for the cell. This causes the oxidative stress and eventually death of the neuronal cell. Also endopeptidase activity is altered by HD. Normally Htt is decomposed by endopeptidases to make sure cleavage products are released. In patients with HD the endopeptidase activity is transformed which causes the cleavage products to aggregate, which again form inclusions [LLBH$^+$02]. These inclusions are characteristic for neurodegenerative diseases such as HD [KMK$^+$14]. The threonine-type (endo)peptidase can not be linked directly to HD, however this is a form of endopeptidase which is altered by mutant HTT.

Clathrin is a protein that ensures a coated vesicle to be able to transport molecules through cells. According to [HW03] are there many proteins which interact with Htt that contribute to clathrin-mediated endocytosis. An mutation in HTT can cause a decrease of this form of endocytose, which might lead to a shortage or surplus of molecules at specific places in or outside of the cell, which can be harmful. Glutamate contributes to the cell death of neurons. According to [RPF11] does the mutant version of HTT among other things lead to cell death of neurons due to a surplus of glutamate. Glutamate is an essential stimulating neurotransmitter, important for brain function and the development of neurons [RPF11]. Another research done by [GPY$^+$85] found out that patients who suffered from HD had a decreased level of glutamate binding. Which causes a surplus of glutamate which again causes the death of neuronal cells.

5.2 Model organisms and human

5.2.1 Proteins

In [Hen23] the five proteins with the highest degree, also called hubs were found to be: ACTB, GADPH, AKT1, TP53 and MYC. To compare these, also was looked at the top-5 hubs in the networks for MMU and RNO. These are visible in Table 10. The top-5 of MMU and RNO were pretty similar, all ten of them are NADH:ubiquinone oxidoreductase subunits [sub]. NADH ubiquinone oxidoreductase is a process which takes place within the mitochondria and it produces ROS, which causes cellular oxidative stress. This form of stress caused by mitochondria is conducing neurodegenerative diseases such as HD [KH06]. So these proteins do play a role in the origin or aggravation of HD.

These ten did not at all match with the top-5 from the HSA network. Four of the five proteins were not even present in the MMU and RNO networks. Only TP53 occurred in the network for rat. In the MMU network TRP53 occurred, this is a ortholog of the human TP53. TP53 is a tumor suppressor protein, it prevents uncontrolled proliferation of cells [Med]. TP53 interacts with Htt. [CBMB05] found out that variations of TP53 can increase or decrease apoptosis and thus play a role in the age of onset of HD.

Also the degrees of the hubs differed much between the organisms. The degrees of the hubs in MMU and RNO were practically equal. However the degrees of the hubs in the HSA network were more than five times as big. The HSA network had 2023 nodes and 82753 edges. Compared to 283 nodes and 6488 edges for RNO and 296 nodes and 6582 edges for MMU. To be able to compare these values the average degree of the five hubs was divided by the total number of edges in the network. This resulted in 0.02 for MMU, 0.02 for RNO and 0.01 for HSA. So the ratio between the

degree of the hubs and the total number of interactions in the network was similar for the three networks. This made that the only difference was between the proteins themselves. Which means there is more difference between human and the model organisms than there is a difference between the two model organisms.

Table 10: Top-5 hubs in the networks. The values represent the degree of the hubs. The hibs in the MMU and RNO column are all NADH:ubiquinone oxidoreductase subunits.

HSA	MMU	RNO
ACTB – 759	Ndufb4 – 156	Ndufb8 – 141
GADPH – 748	Ndufb8 – 141	Ndufb7 – 139
AKT1 – 715	Ndufa9 – 139	Ndufb9 – 137
TP53 – 678	Ndufb7 – 138	Ndufa9 – 137
MYC – 578	Ndufv1 – 137	Ndufa10 – 137

5.2.2 Functional enrichment

As seen, the 6 GO Molecular Functions and GO Biological Processes of MMU and RNO were the same, however this was not the case when comparing them to the HSA network. Table 11 shows the five MFs and BPs of the HSA network. Comparing them to the ones found in the networks of MMU and RNO visible in Table 8 and Table 7 for the GO Biological Processes and GO Molecular Functions respectively, showed that these were very different. Not one of the Molecular Functions and Biological Processes that were in the tables of the model organisms were found in the human consensus network.

Table 11: Ids and GO terms of the GO Molecular Functions and GO Biological Processes according to Cytoscape of the human consensus network made by [Hen23].

GO Molecular Function	GO Biological Process
Protein binding (GO:0005515)	Regulation of biological quality (GO:0065008)
Binding (GO:0005488)	Response to stimulus (GO:00550896)
Enzyme binding (GO:0019899)	Response to chemical (GO:0042221)
Identical protein binding (GO:0042802)	Response to organic substance (GO:0010033)
Signaling receptor binding (GO:0005102)	Positive regulation of biological process (GO:0048518)

To be able to really compare the results of the networks for MMU and RNO to human, the clusters of interest chosen and made by [Hen23] were analyzed using clusterProfiler. The results are visible in Figure 10. When comparing this image to the ones of MMU and RNO some overlap was seen, unlike the results obtained with Cytoscape. Below the five enriched terms which occur in the human clusters which also occur in the clusters for MMU and RNO are visible. The fact that there is overlap between the three organisms means that part of the functions and processes of the proteins are similar. All of these five overlapping enriched terms between the three organisms are already described and linked to HD in section 5.1.2. Because of the fact that there were some extra clusters obtained from the MMU network, was looked whether or not they could be linked to the HSA network. Unfortunately this was not the case.

- RNO/MMU-HSA
  - GO:0009055 – electron transfer activity
  - GO:0003954 – NADH dehydrogenase activity
  - GO:0008137 – NADH dehydrogenase (ubiquinone) activity
  - GO:0008066 – glutamate receptor activity
  - GO:0030276 – clathrin binding

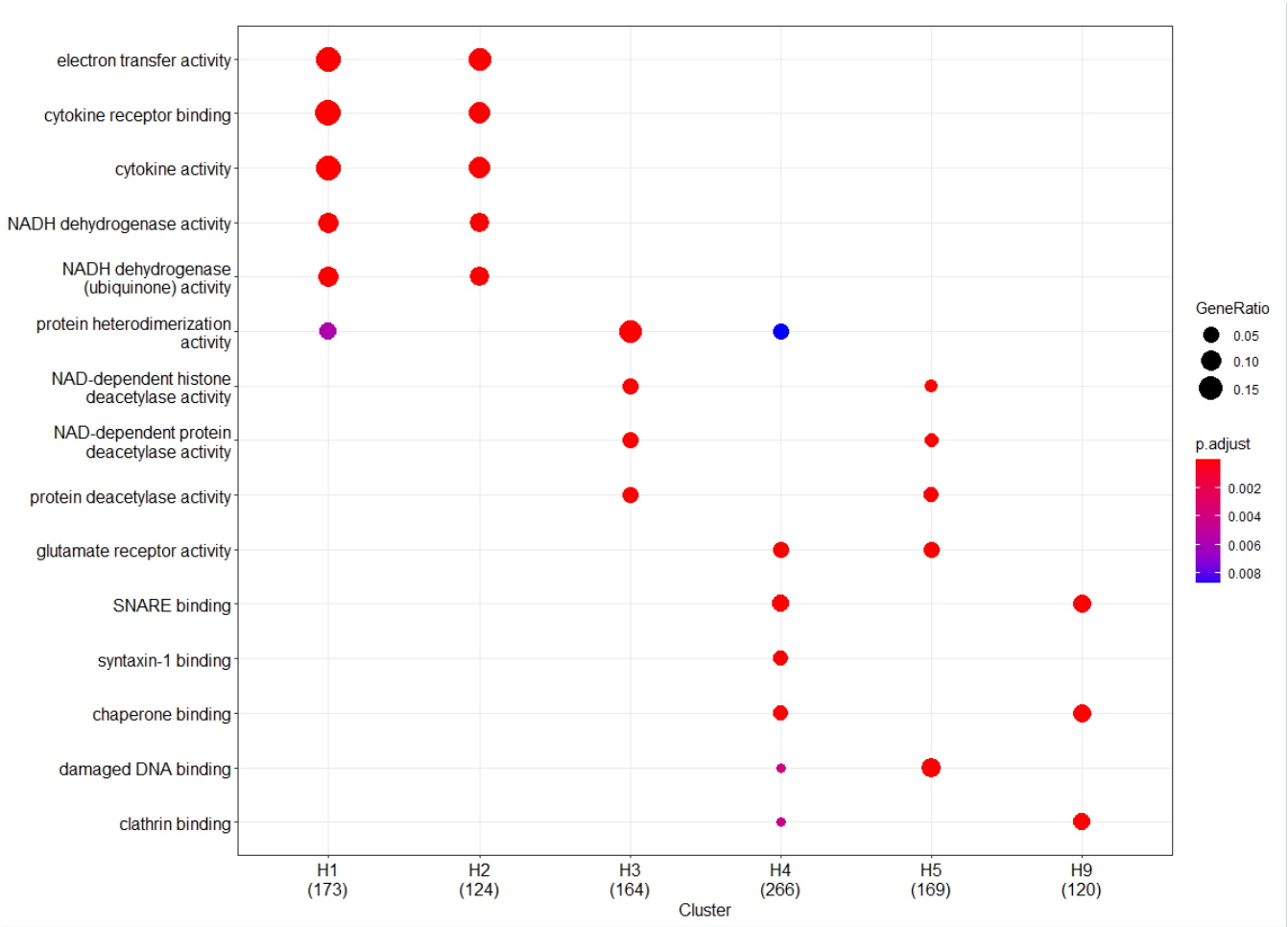

Figure 10: Top-3 enriched terms per cluster of interest for HSA. Clusters are made by [Hen23]. Syntaxin-1 binding is the only enriched term connected to one cluster, which is cluster 4.

6 Conclusions and Further Research

This thesis resulted in protein-protein interaction networks regarding Huntington's disease in the model organisms MMU and RNO. Mice and rats are commonly used as model organisms for Huntington's disease. The sequences of the genes in the three organisms do not differ much, but

whether or not the disease mechanism in these model organisms is similar to the human one is not clear. Protein-protein interaction networks are created to investigate the interacting proteins and to be able to compare the different organisms and thereby look at possible similarities and differences. The networks were made using Cytoscape and analyzed using Cytoscape and the R package clusterProfiler to be able to compare the networks. The comparison was not only done between the two model organisms, also the created networks were compared to the HSA HD network made by [Hen23].

According to the research done in this thesis can be concluded that the networks of MMU and RNO do not differ much at all. The networks of these two model organisms are very similar, they differ in appearance but have the same functions and also the enriched terms of the networks in both organisms are the same. When comparing the created networks to the HSA network, way more differences are visible. The HSA network is bigger, however there are enriched terms which overlap between human and the model organisms. This is exactly what would be expected when comparing different organisms. A different organism means a different gene sequence which means different proteins. The fact that these different proteins have some of the same enriched terms is a positive thing. Now can be inferred that the proteins in the three different organisms have some overlap in function, and may have a similar disease method.

To come back on the research question, yes there are both differences and similarities when comparing model organisms to human. Looking at the proteins in the network there is not much overlap between humans and the model organisms. However, when looking at the enriched terms of the three networks, there is some overlap. Which means that the proteins differ in the three organisms but partly have the same function and the three organisms share common core network features. Which is exactly what would be expected when comparing PPINs of different organisms.

The lack of overlap can be caused by the difference in size of the HSA network compared to the networks for MMU and RNO. The HSA network contains over a thousand proteins and over 80000 interactions between these proteins. The networks for MMU and RNO contain only 300 proteins and 6500 interactions. The sequences of the three HTT genes in the organisms are quite similar, however, the fact that there is not much overlap when looking just at the proteins in the networks may be caused by HTT gene having different interacting proteins in different organisms. The fact that the clusters in MMU and RNO that contain Htt, are almost equal and thus have the same Htt-interacting proteins, may lead to a similar disease mechanism in these model organisms. However the differences in Htt-interacting proteins between human and the model organisms are likely to cause a different disease mechanism when comparing human to these model organisms.

Further research could try to expand the networks of MMU and RNO, trying to find more proteins connected to Huntington's Disease in these model organisms. For example looking at more pathway databases. When the networks are bigger a better comparison can be done to the HSA network. The networks can also be improved by creating a first and second neighbor network of the model organisms. Also this research can be expanded by adding expression data to look for up- and down-regulated proteins in the network, which again can be compared between the three networks. These two expansions might lead to more important similarities or differences found between the HSA network and the networks of the model organisms.

References

[ANK+13] Frederik Heurlin Aidt, Signe Marie Borch Nielsen, Jørgen Kanters, Dominik Pesta, Troels Tolstrup Nielsen, Anne Nørremølle, Lis Hasholt, Michael Christiansen, and Christian Munch Hagen. Dysfunctional mitochondrial respiration in the striatum of the huntington's disease transgenic r6/2 mouse model. *PLoS currents*, 2013.

[AZX+22] Fengmao An, Ruyi Zhao, Xinran Xuan, Tianqi Xuan, Guowei Zhang, and Chengxi Wei. Calycosin ameliorates advanced glycation end product-induced neurodegenerative changes in cellular and rat models of diabetes-related alzheimer's disease. *Chemico-Biological Interactions*, 368:110206, 2022.

[BDG+15] Gillian P. Bates, Ray Dorsey, James F. Gusella, Michael R. Hayden, Chris Kay, Blair R. Leavitt, Martha Nance, Christopher A. Ross, Rachael I. Scahill, Ronald Wetzel, Edward J. Wild, and Sarah J. Tabrizi. Huntington disease. *Nature reviews disease primers*, 1(1), 2015.

[BH03] Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4, 2003.

[CAS+] Rajnish K. Chaturvedi, Peter Adhihetty, Shubha Shukla, Thomas Hennessy, Noel Calingasan, Lichuan Yang, Anatoly Starkov, Mahmoud Kiaei, Milena Cannella, Jenny Sassone, Andrea Ciammola, Fernando Squitieri, and M. Flint Beal. Impaired pgc-1 function in muscle in huntington's disease. *Human molecular genetics*, 18(16):3048–3065.

[CBMB05] Biswanath Chattopadhyay, Kanad Baksi, Saikat Mukhopadhyay, and Nitai P. Bhattacharyya. Modulation of age at onset of huntington disease patients by variations in tp53 and human caspase activated dnase (hcad) genes. *Neuroscience Letters*, 374(2):81–86, 2005.

[Con] Cytoscape Consortium. stringapp.

[Con18] Cytoscape Consortium. What is cytoscape?, 2018.

[Con23] GO Consortium. The gene ontology research, 2023.

[CTG96] O Coux, K Tanaka, and A L Goldberg. Structure and functions of the 20s and 26s proteasomes. *Annual review of biochemistry*, 1996.

[DB18] Stephen B Dunnett and Simon P Brooks. Motor assessment in huntington's disease mice. *Methods Mol Biol.*, 2018.

[dB23] Aster de Boer. Identifying overlapping processes of alzheimer's and huntington's disease with a protein-protein interaction network analysis. *Thesis Bachelor Bioinformatica*, 2023.

[EBPH09] Dagmar E. Ehrnhoefer, Stefanie L. Butland, Mahmoud A. Pouladi, and Michael R. Hayden. Mouse models of huntington disease: variations on a theme. *Disease Models Mechanisms*, 2(3-4):123–129, 2009.

[EE02] EMBL-EBI. Go:003954 nadh dehydrogenase activity, 2002.

[EE08a] EMBL-EBI. Go:0006091 generation of precursor metabolites and energy, 2008.

[EE08b] EMBL-EBI. Go:0006119 oxidative phosphorylation, 2008.

[EE08c] EMBL-EBI. Go:0016491 oxidoreductase activity, 2008.

[EE08d] EMBL-EBI. Go:0046034 atp metabolic process, 2008.

[EE09] EMBL-EBI. Go:0003824 catalytic activity, 2009.

[EE10] EMBL-EBI. Go:0022904 respiratory electron transport chain, 2010.

[EE14] EMBL-EBI. Go:0045333 cellular respiration, 2014.

[EE17] EMBL-EBI. Go:0009055 electron transfer activity, 2017.

[EE18] EMBL-EBI. Go:0015078 proton transmembrane transporter activity, 2018.

[EE21] EMBL-EBI. Go:0022900 electron transport chain, 2021.

[EE22] EMBL-EBI. Go:008137 nadh dehydrogenase (ubiquinone) activity, 2022.

[EKGB16] Scott Emmons, Stephen Kobourov, Mike Gallant, and Katy Börner. Analysis of network clustering algorithms and cluster quality metrics at scale, 2016.

[GGM+96] M. Gu, M. T. Gash, V. M. Mann, F. Javoy-Agid, J. M. Cooper, and A. H. V. Schapira. Mitochondrial defect in huntington's disease caudate nucleus. *Annals of Neurology*, 39(3):385–389, 1996.

[GMLDVG+22] Adrian Garcia-Moreno, Raul López-Domínguez, Juan Antonio Villatoro-García, Alberto Ramirez-Mena, Ernesto Aparicio-Puerta, Michael Hackenberg, Alberto Pascual-Montano, and Pedro Carmona-Saez. Functional enrichment analysis of regulatory elements. *Biomedicines*, 10, 2022.

[GPY+85] J. Timothy Greenamyre, John B. Penney, Anne B. Young, Constance J. D'Amato, Samuel P. Hicks, and Ira Shoulson. Alterations in l-glutamate binding in alzheimer's and huntington's diseases. *Science*, 227(4693):1496–1499, 1985.

[Hen23] Nina Henninger. Extending consensus knowledge in huntington's disease protein interaction networks. 2023.

[HF12] Ainhi D. Ha and Victor S.C. Fung. Huntington's disease. *Current Opinion in Neurology*, 25(4):491–498, 2012.

[HW03] Phoebe Harjes and Erich E Wanker. The hunt for huntingtin function: interaction partners tell many different stories. *Trends in Biochemical Sciences*, 28(8):425–433, 2003.

[JB12] Ashu Johri and M. Flint Beal. Antioxidants in huntington's disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1822(5):664–674, 2012.

[Jia22] Chen Ji Rong Jiang. Finding consensus knowledge in the huntington's disease pathway. *Thesis Bachelor Bioinformatica*, 2022.

[KH06] Lothar Kussmaul and Judy Hirst. The mechanism of superoxide production by nadh:ubiquinone oxidoreductase (complex i) from bovine heart mitochondria. *Proceedings of the national academy of sciences*, 102(20):7607–7612, 2006.

[KMK+14] Masaru Kurosawa, Gen Matsumoto, Yoshihiro Kino, Misako Okuno, Mizuki Kurosawa-Yamada, Chika Washizu, Harumi Taniguchi, Kazuhiro Nakaso, Toru Yanagawa, Eiji Warabi, Tomomi Shimogori, Takashi Sakurai, Nobutaka Hattori, and Nobuyuki Nukina. Depletion of p62 reduces nuclear inclusions and paradoxically ameliorates disease phenotypes in Huntington's model mice. *Human Molecular Genetics*, 24(4):1092–1105, 2014.

[Lab] Kanehisa Laboratories. Kegg: Kyoto encyclopedia of genes and genomes.

[LLBH+02] Astrid Lunkes, Katrin S. Lindenberg, Léa Ben-Haıem, Chantal Weber, Didier Devys, G.Bernhard Landwehrmeyer, Jean-Louis Mandel, and Yvon Trottier. Proteases acting on mutant huntingtin generate cleaved products that differentially build up cytoplasmic and nuclear inclusions. *Molecular Cell*, 10(2):259–269, 2002.

[Med] MedlinePlus. Tp53gene.

[Nat] Nature. Protein–protein interaction networks articles from across nature portfolio.

[OS01] M. Orth and A.H.V. Schapira. Mitochondria and degenerative disorders. *American Journal of Medical Genetics*, 106(1):27–36, 2001.

[PSSPC+04] Francisca Pérez-Severiano, Abel Santamaría, José Pedraza-Chaverri, Omar N. Medina-Campos, Camilo Ríos, and José Segovia. Increased formation of reactive oxygen species, but no changes in glutathione peroxidase activity, in striata of mice transgenic for the huntington's disease mutation. *Neurochemical Research*, 29:729–733, 2004.

[RHT] Paul D. Ray, Bo-Wen Huang, and Yoshiaki Tsuji. Reactive oxygen species (ros) homeostasis and redox regulation in cellular signaling. *Cellular Signaling*, 24(5):981–990.

[RPF11] Fabiola M. Ribeiro, Rita G.W. Pires, and Stephen S. G. Ferguson. Huntington's disease and group i metabotropic glutamate receptors. *Molecular Neurobiology*, 43, 2011.

[SL11] J. Schulte and J. T. Littleton. The biological function of the huntingtin protein and its relevance to huntington's disease pathology. *Current trends in neurology*, 5:65–78, 2011.

[SR03] Katharine L. Sugars and David C. Rubinsztein. Transcriptional abnormalities in huntington disease. *Trends in Genetics*, 19(5):233–238, 2003.

[sub] Nadh:ubiquinone oxidoreductase supernumerary subunits (nduf).

[SZK+15] Levente Szalárdy, Dénes Zádori, Péter Klivényi, József Toldi, and László Vécsei. Electron transport disturbances and neurodegeneration: From albert szent-györgyi's concept (szeged) till novel approaches to boost mitochondrial bioenergetics. *Oxidative medicine and cellular longevity*, 2015.

[Szk21] Damian Szklarczyk. The string database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, 49:605–612, 2021.

[vDvdVRB86] J. G. van Dijk, E. A. van der Velde, R. A. C. Roos, and G. W. Bruyn. Juvenile huntington disease. *Human genetics*, 73(3):235–239, 1986.

[VFMV03] Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature Biology*, 21:697–700, 2003.

[Wal07] Francis O Walker. Huntington's disease. *The Lancet*, 369(9557):218–228, 2007.

[WHX+21] Tianzhi Wu, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Lang Zhou, Wenli Tang, Li Zhan, Xiaocong Fu, Shanshan Liu, Xiaochen Bo, and Guangchuang Yu. clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141, 2021.

[WLDP10] Jianxin Wang, Min Li, Youping Deng, and Yi Pan. Recent advances in clustering methods for protein interaction networks. *BMC Genomics*, 11(Suppl 3), 2010.

[YWHH12] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, 2012.

A MMU

Below the networks for MMU can be found.

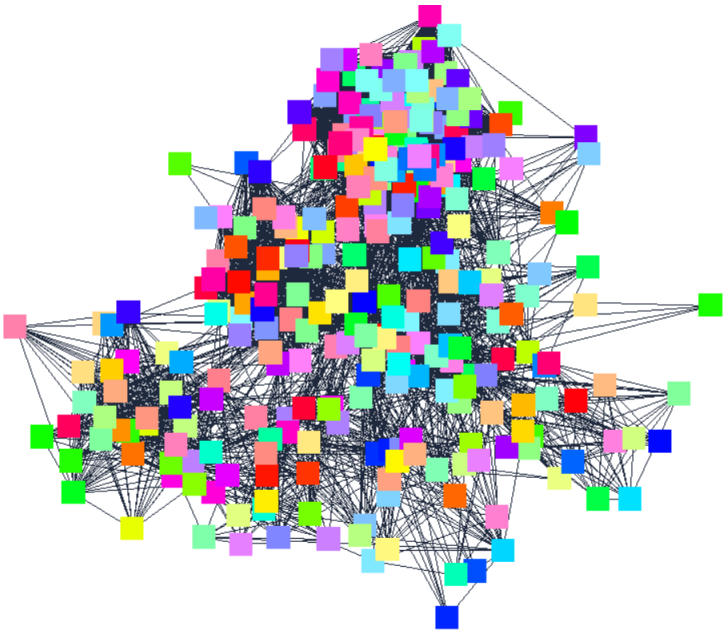

Figure 11: MMU STRING network with 0.4 cut-off score.

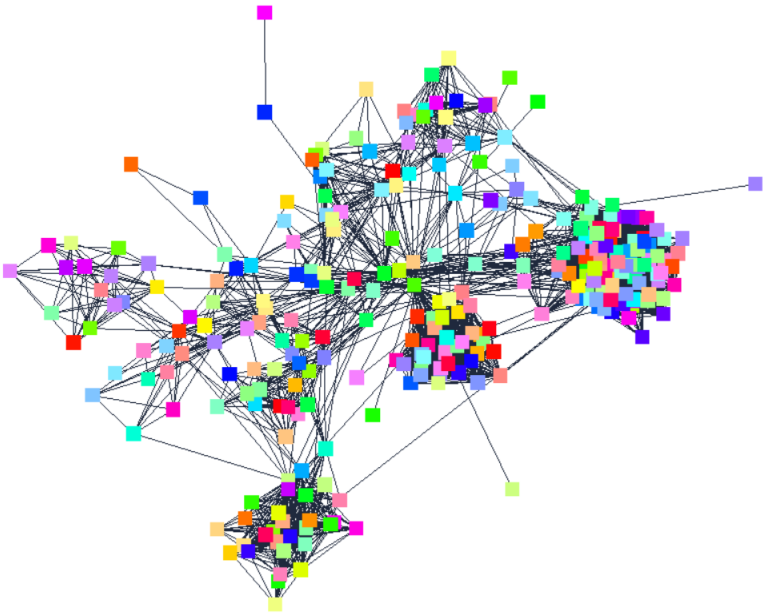

Figure 12: MMU STRING network with 0.7 cut-off score.

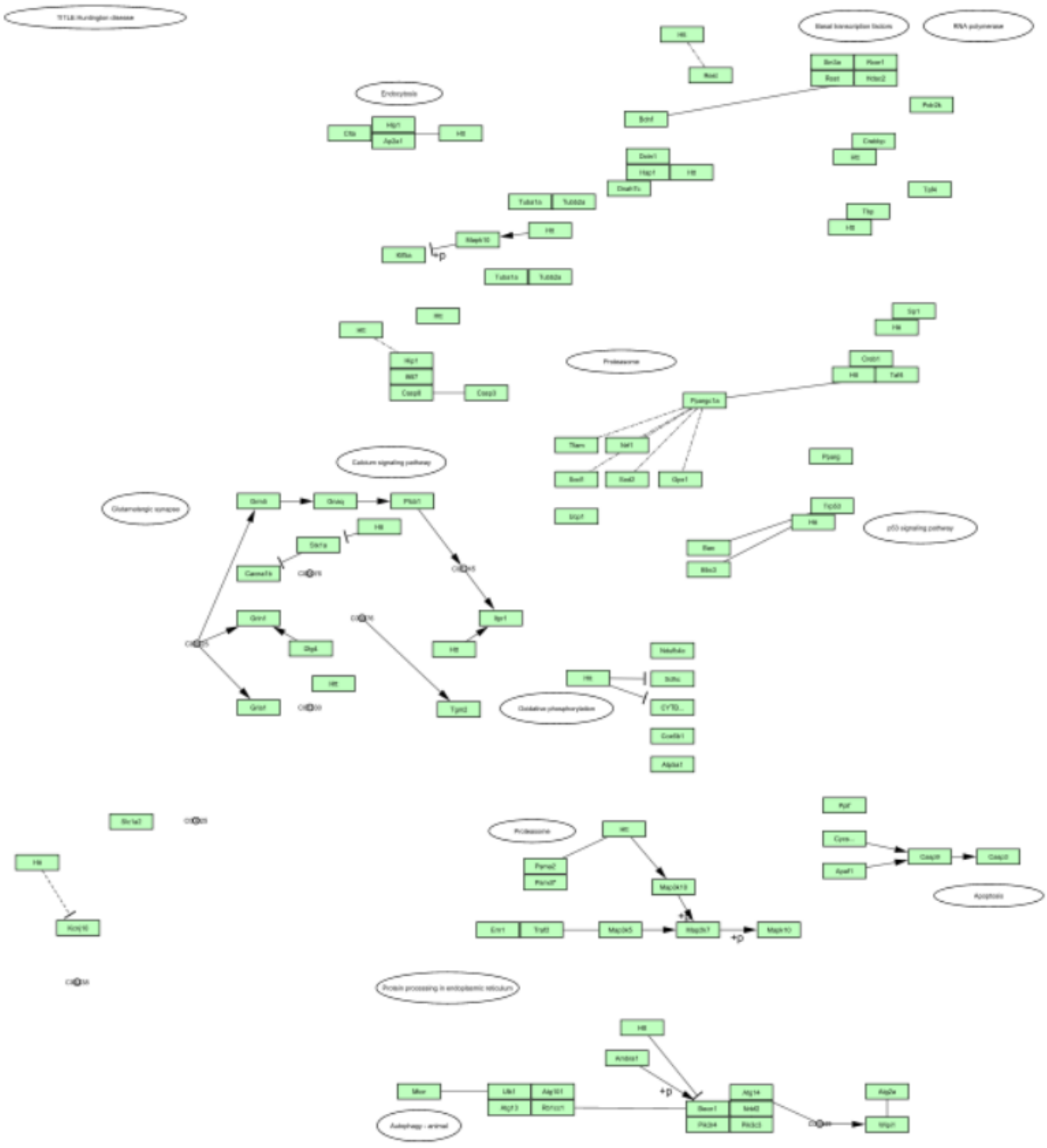

Figure 13: MMU KEGG pathway for Huntington's disease.

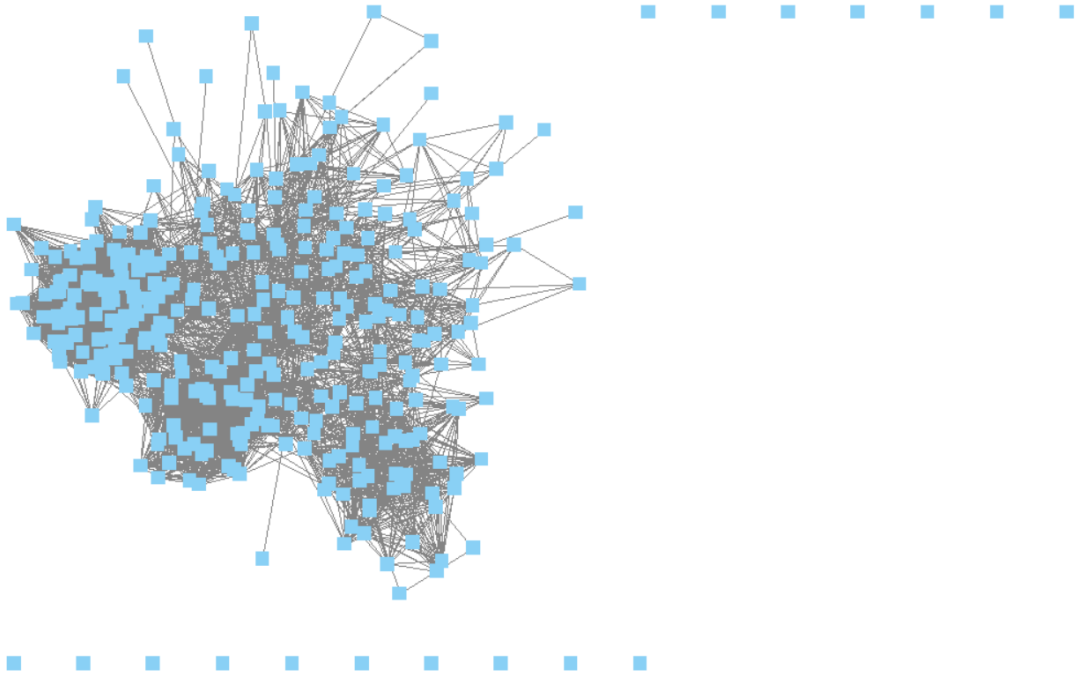

Figure 14: MMU STRING network with 0.4 cut-off score merged with the KEGG pathway.

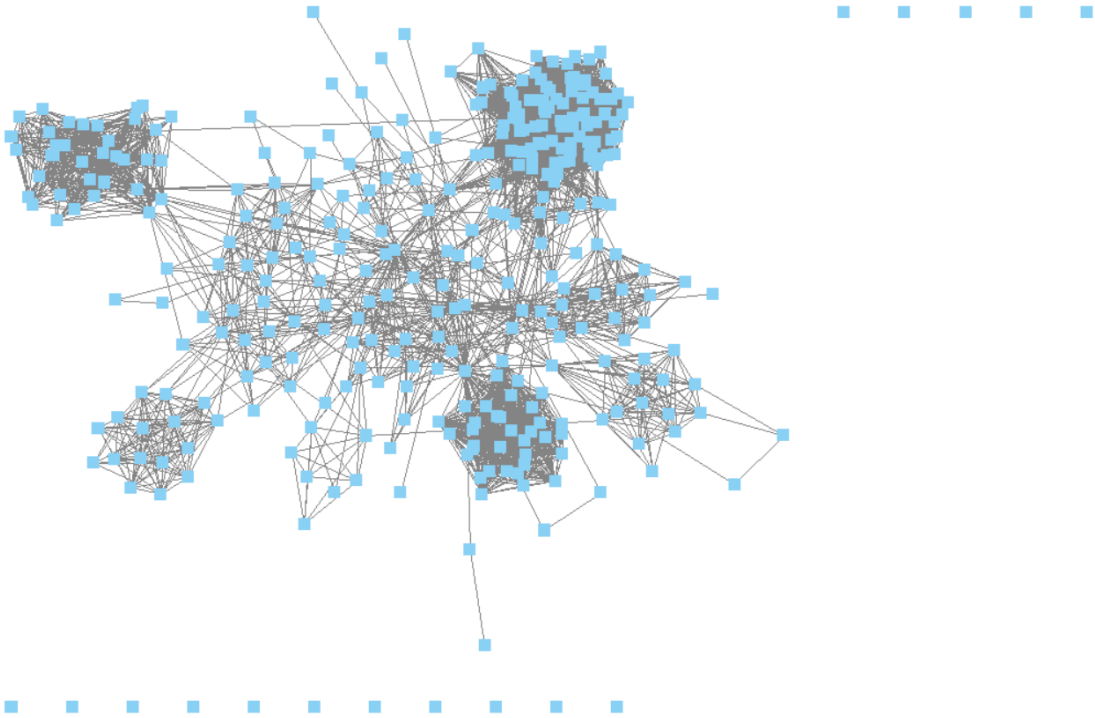

Figure 15: MMU STRING network with 0.7 cut-off score merged with the KEGG pathway

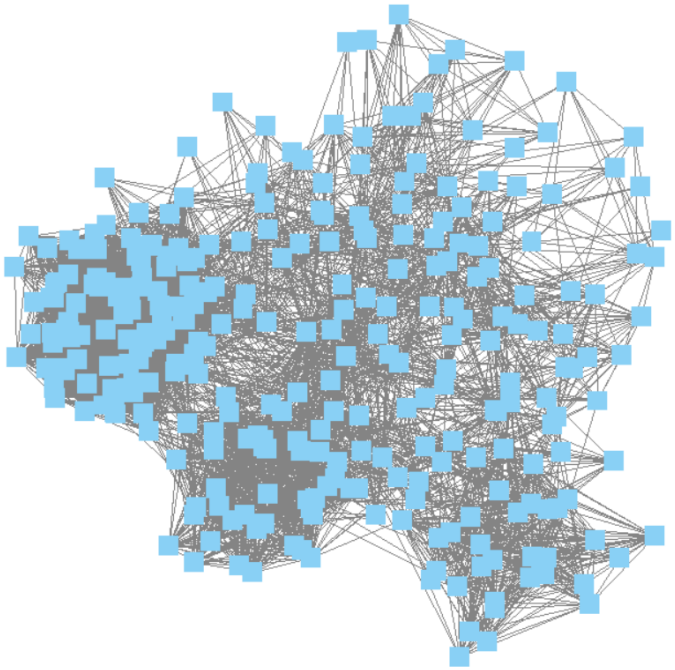

Figure 16: MMU STRING network with 0.4 cut-off score merged with the KEGG pathway and filtered on nervous system.

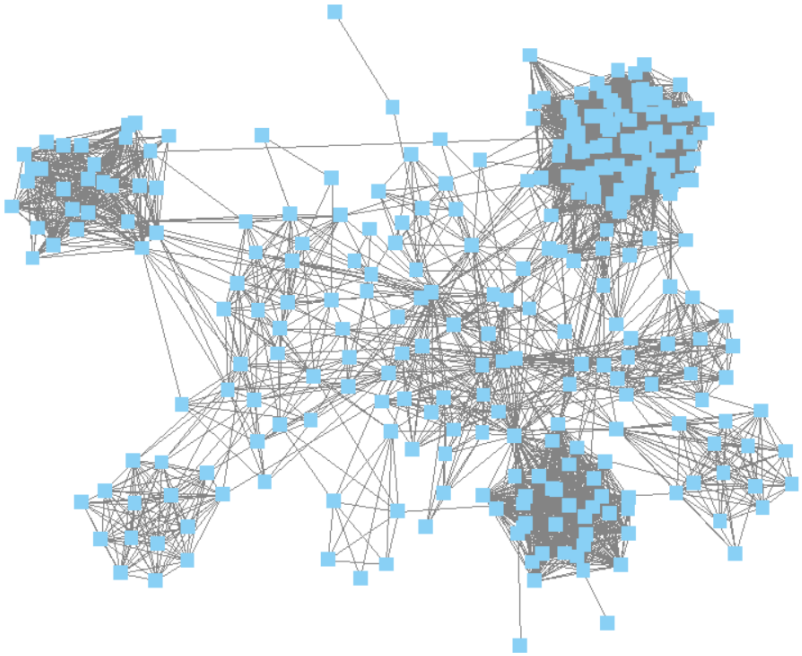

Figure 17: MMU STRING network with 0.7 cut-off score merged with the KEGG pathway and filtered on nervous system.

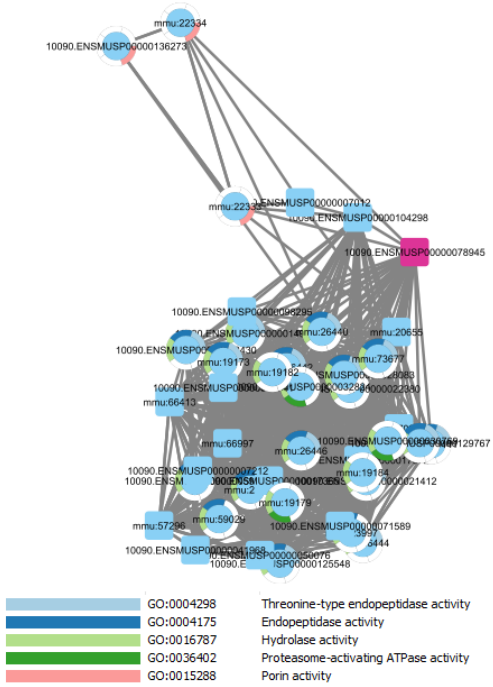

Figure 18: GO Molecular Function enrichment for cluster 2 of MMU, containing Htt which is colored purple.

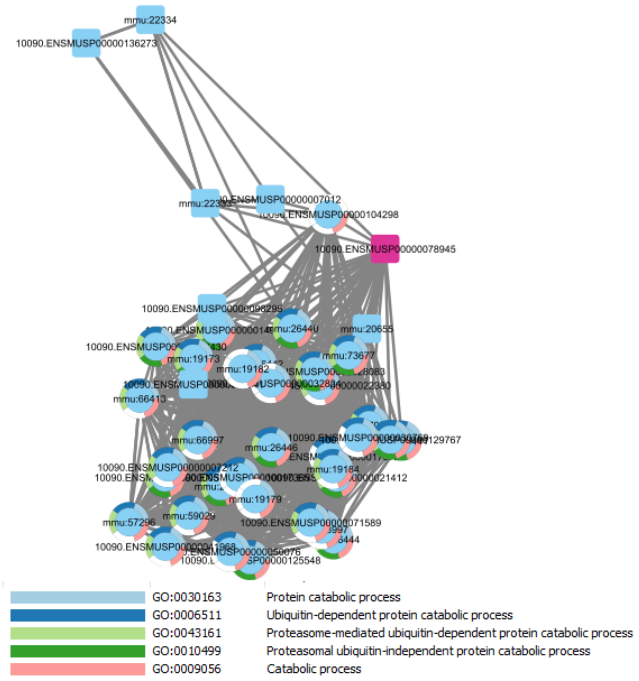

Figure 19: GO Biological Process enrichment for cluster 2 of MMU, containing Htt which is colored purple.

B RNO

Below the networks for RNO can be found

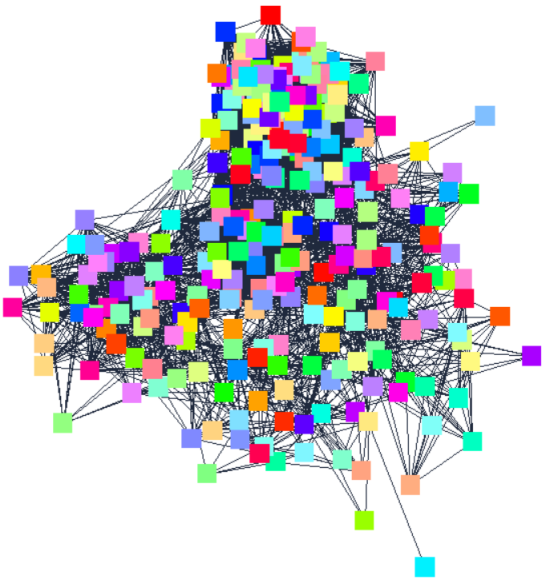

Figure 20: RNO STRING network with 0.4 cut-off score.

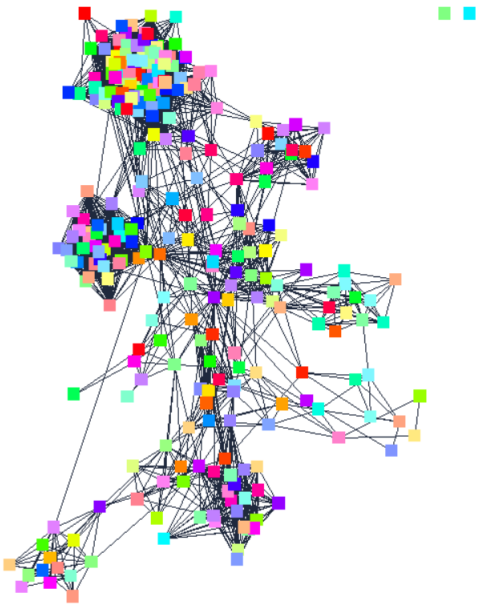

Figure 21: RNO STRING network with 0.7 cut-off score.

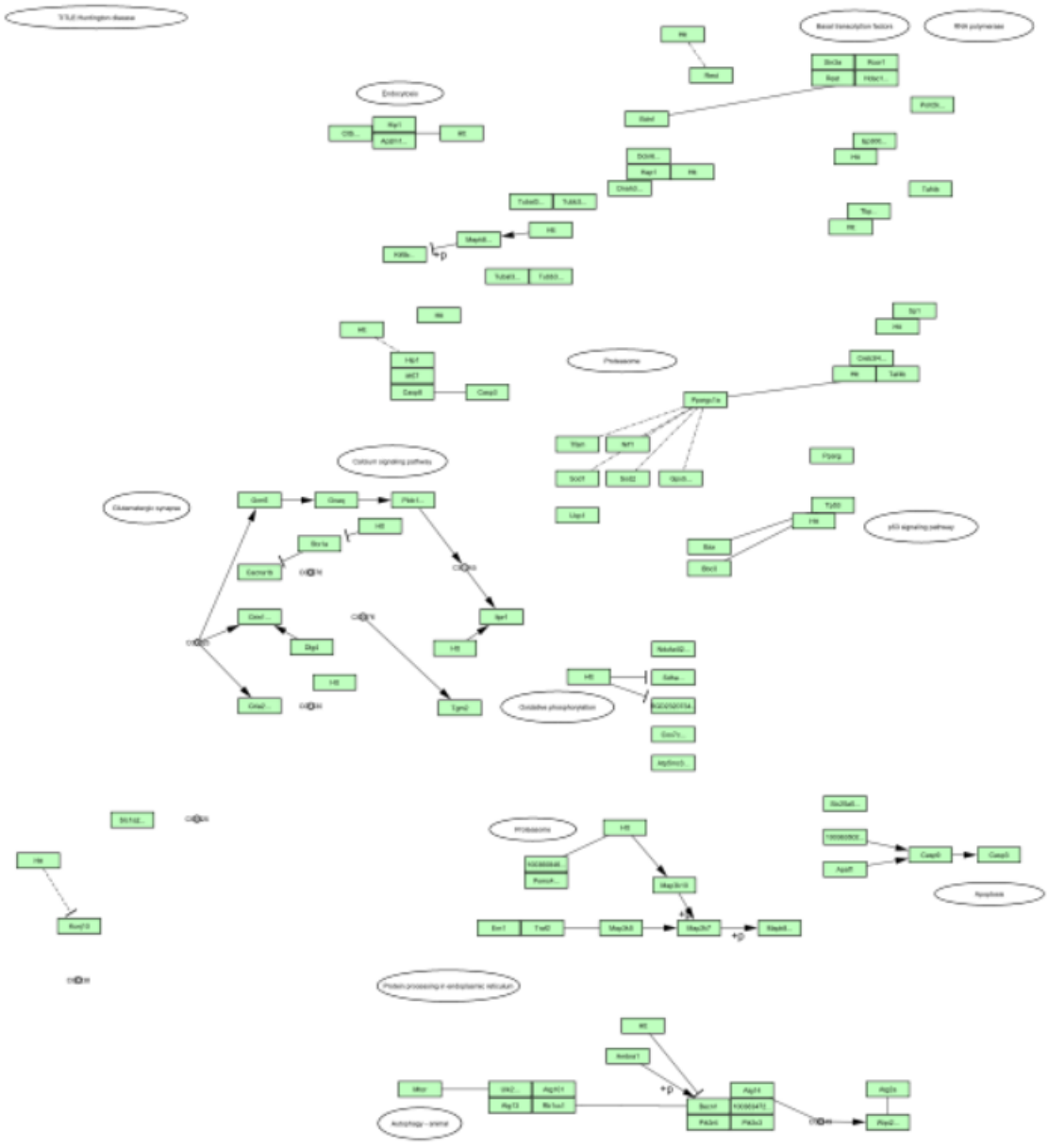

Figure 22: RNO KEGG pathway for Huntington's disease.

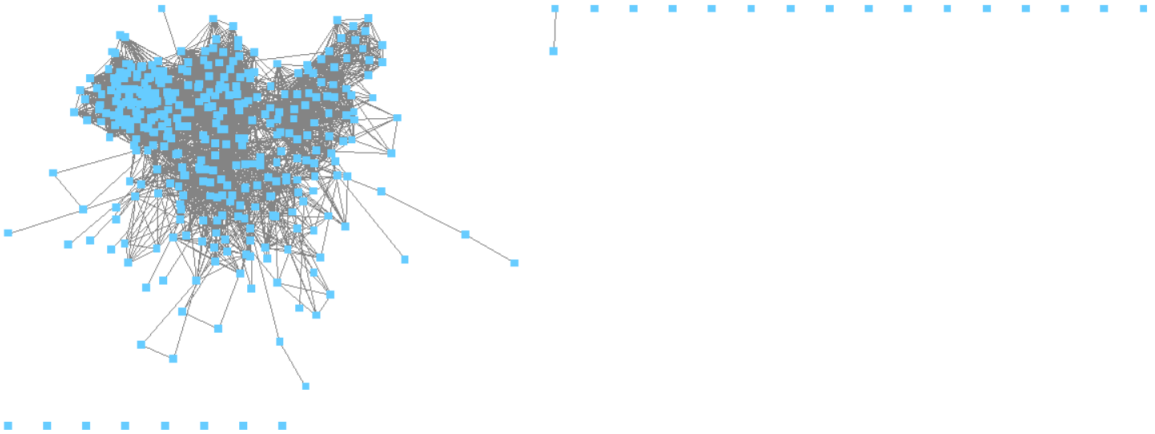

Figure 23: RNO STRING network with 0.4 cut-off score merged with the KEGG pathway.

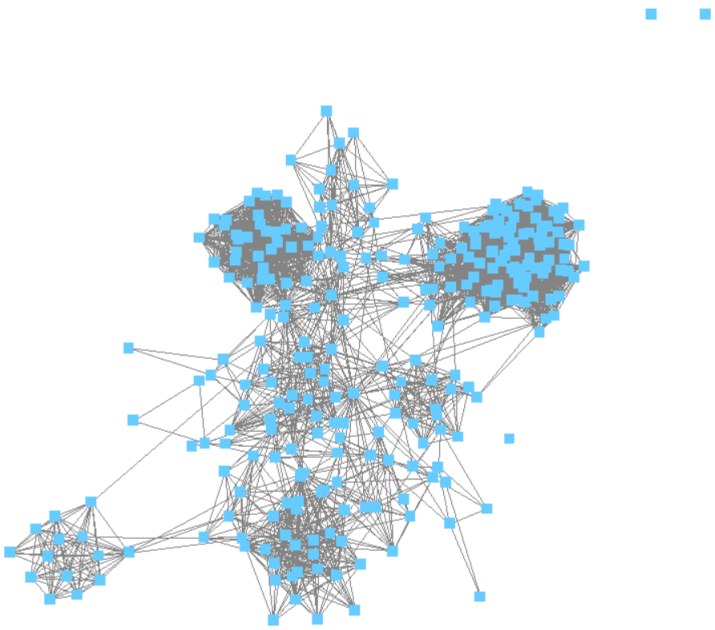

Figure 24: RNO STRING network with 0.7 cut-off score merged with the KEGG pathway

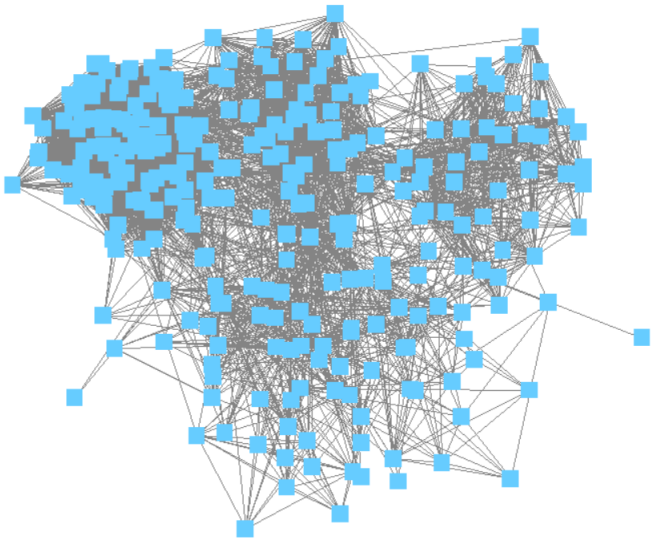

Figure 25: RNO STRING network with 0.4 cut-off score merged with the KEGG pathway and filtered on nervous system.

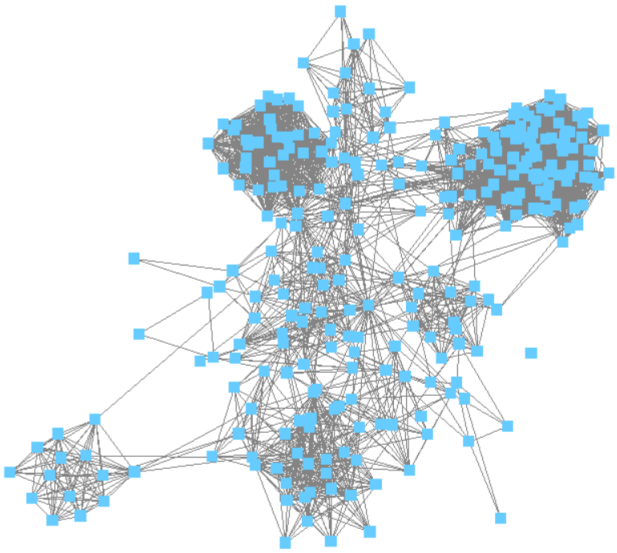

Figure 26: RNO STRING network with 0.7 cut-off score merged with the KEGG pathway and filtered on nervous system.

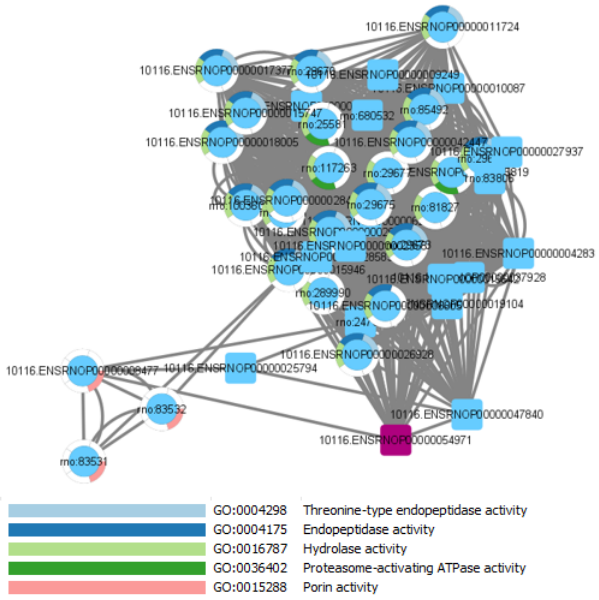

Figure 27: GO Molecular Function enrichment for cluster 2 for RNO. The cluster contains Htt which is colored purple.

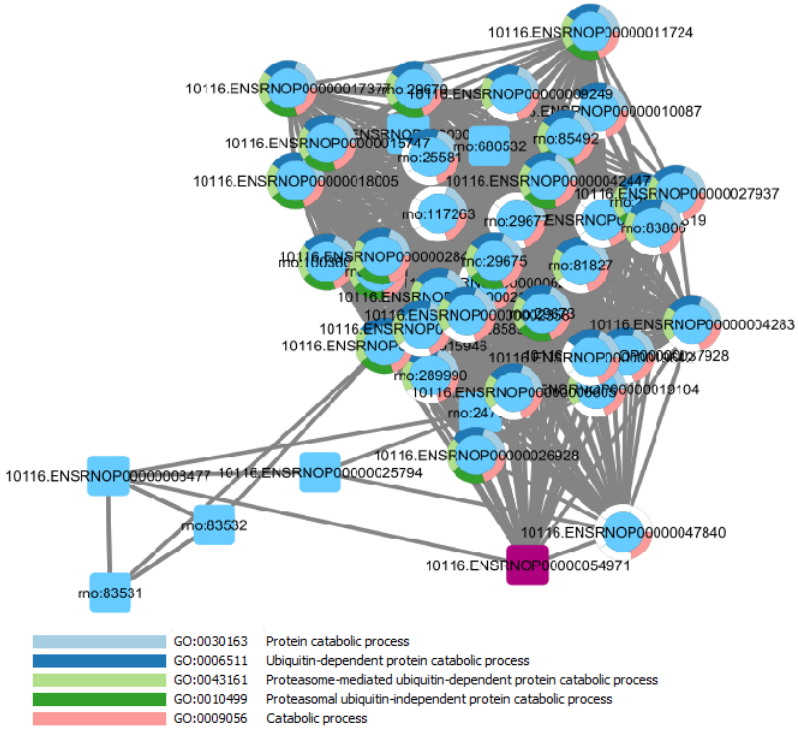

Figure 28: GO Biological Process enrichment for cluster 2 for RNO. The cluster contains Htt which is colored purple.