Universiteit Leiden
The Netherlands

Opleiding Informatica

Revisiting the evolution and function of Polyglutamine Repeats

Lucas Heijnen

Supervisors:
Dr. K. J. Wolstencroft
Dr. A. P. Goultiaev

BACHELOR THESIS

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

09/01/2020

Abstract

Polyglutamine (PolyQ) proteins are proteins with polyglutamine tracts of 10 to several hundred glutamines. When these polyQ tracts exceed a certain normal range through mutations, some of these proteins cause neurodegenerative diseases. In 2012, Schaefer *et al* concluded in their research that focus on polyQ proteins should shift focus from disease causing proteins to wild type proteins, this thesis aims to recreate the research done in 2012 with novel data. Results show that the fraction of polyQ proteins have not changed relatively to the fractions found in 2012. The functions of domains overrepresented in polyQ proteins in 2012 are almost all found in the novel results, however, many new functions have been found, mostly in DNA and RNA binding. The protein functions of human polyQ proteins are in line with the domain functions, being mostly functional in DNA and RNA binding.

Contents

Chapter 1

Introduction

Protein domains are parts of proteins which are conserved through different species and have several functions, these functions are also conserved [1]. These domains can fold themselves and function independently, because of this, domains can be seen as building blocks for proteins when the sequence for a specific domain is inserted in a protein through evolution. This sequence will fold itself into the domain to function within the protein without need for folding proteins etc. Some examples of functions of domains are binding domains, which bind to specific molecules or atoms, and transmembrane domains, which is a segment of a protein that spans a membrane. A protein domain can often stabilize itself and folds independently from the rest of the protein. Most proteins are made up of several protein domains which together form a new function than the separate domains would form. Domains are folded in such ways that they can be placed into proteins during evolution, possibly creating new proteins with new functions. Most protein domains have considerable amounts of functions and structures, like the zinc finger domain [2]. For most domains, the function is know, but there still are a lot of domains with functions unknown, in the current Pfam release (Pfam 32.0), 22% of the entries are "domains of unknown functions"(DUFs) [4]. It is known that mutated protein domains can be pathogenic when these mutations prevents the domain from functioning properly [5] [6] [7].

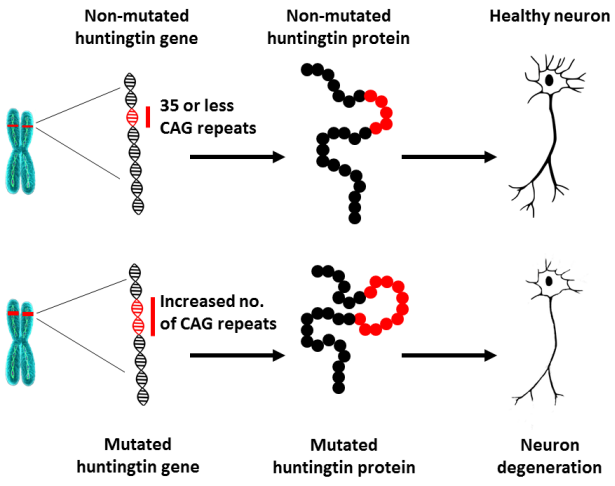

Figure 1.1: The cause of Huntington's disease [3]

Polyglutamine (polyQ) tracts are a natural part of proteins that occur in several different species. Long chains of glutamine residues have been found to be pathogenic in humans. In at least nine proteins, an expanded polyQ tract results in a neurodegenerative disease, most common is Huntington's disease [8]. These longer polyQ tracts often are misfolded and aggregate to insoluble protein

clumps [9], which interfere with neuron functions [10].
When these polyQ tracts are not elongated in polyQ proteins, these proteins have normal functions, like stabilizing interactions between proteins [11]. It is most likely that this stabilization happens through structural changes where the polyQ region interacts with a coiled-coil region to make this region more suitable for interaction with a coiled-coil region of another protein. When the polyQ region is elongated, it is thought that the proteins would interact differently with proteins than normal polyQ regions, which then causes protein aggregation.

1.1 Research goal

This research is a continuation of a research done in 2012 by Schaefer *et al* [11]. This research was aimed at analyzing information available in online databases of polyQ tracts to research the function of these tracts. Schaefer concluded that the research of polyQ proteins and the expansion of these regions should be shifted towards wild-type polyQ proteins, as a better understanding of the influence of these proteins could also give more insight into the neurodegenerative diseases originating from expanded polyQ tracts.
Since 2012, a huge amount of DNA has been sequenced in many different organisms, including a lot of new polyQ proteins. In this research, our main goal is to study the function of polyglutamine repeats with a similar but improved approach of Schaefer *et al*, with a number of other organisms and many more polyglutamine proteins. Doing this research using more data, we hope to find new insights into possible functions of polyQ regions and proteins. This may provide more insight in the biological processes that are disrupted in diseases associated with elongated polyQ repeats.

1.2 Related work

1.2.1 PolyQ domain

The importance of protein domains for our understanding of protein function and structure has led to several automated domain identification tools. Domain identification involves pattern matching in the sequence data against known domains. Polyglutamine proteins have long been studied because of their disease-causing mutation, the expansion of the polyglutamine tract.
A polyglutamine tract is a part of a protein which consists of several glutamine amino acids, which are coded by codons CAA and CAG [12]. Normally, these tracts are between 10 and 100 amino acids long, but when expanded, can increase considerably, for example with the polyQ protein Huntingtin, normal lengths of the polyQ region in this protein are between seven and 35, but the highest reported length was around 250 amino acids [13].
The function of these proteins is to guide other protein-protein interactions. However, for a few polyQ proteins, when the polyQ region is expanded, these proteins affect other protein-protein interactions. It is thought these

"wrong" interactions cause protein aggregation, which leads to the several different diseases [11].
The polyQ 2.0 database [14] was described as a extensive database for human polyQ proteins with various different annotations to examine these proteins in greater detail. Of course, having only human polyQ proteins, this database can not be used as a complete dataset for this research, but its structure can be used to create our own database.
As mentioned before, we will recreate experiments done by Schaefer *et al* in which they systematically analyzed data taken from sequence databases to find novel functions and interactions of polyQ proteins in several organisms. Schaefer looked at different properties of polyQ proteins, like their function, emergence in protein families and proteins interacting with polyQ proteins. The results suggest polyQ proteins having a role in protein interactions, specifically with coiled-coil regions.

1.2.2 Domain presence to find protein functions

Proteins are mostly made up of domains, these domains can be seen as building blocks for proteins with seperate functions. Most of the protein sequence that is relevant to protein function is made up of these domains [15]. This implies that the protein function can largely be predicted only looking at the domains found in the protein. Looking at which domains are found together in proteins can also help to find the protein function, as it is known that several domains are often found together, which could indicate these domains combined create new functions.

1.3 Overview

This research project is a bachelor thesis for the bachelor Bioinformatics at Leiden university, supervised by Dr. K. J. Wolstencroft.
Chapter 2 of this research shows all methods used to collect and refine the data and all methods to examine the data found. Chapter 3 shows the results found in this research. Chapter 4 concludes the findings.

Chapter 2

Materials & Methods

2.1 PolyQ data set definition

To be able to compare the results found with the results of Schaefer *et al* [11], we used similar restrictions of the polyQ data set.
For a protein to be a polyQ protein, we use the threshold of 10 consecutive glutamines, Scheafer *et al* found this to be the minimum length where all known human polyQ diseases were recognized [8]. To find all proteins with a polyQ tract of at least 10 consecutive Qs, Uniprot peptide search was used [16]. Uniprot is a freely accessible protein database containing proteins from most known organisms. We used both manually curated Swiss-Prot entries and automatically curated TrEMBL entries, because over the last few years, TrEMBL entries have grown at such a speed that they cannot be manually curated, so using only Swiss-Prot entries would result in incomplete data. The Uniprot peptide search returned over 144.000 polyQ proteins. We then defined the set of species used in our analyses as all species with eight or more polyQ proteins, at least 750 protein entries in Uniprot and with fully sequenced genomes [17], the same threshold as in 2012. We also removed several organism strains, choosing only the most common one per organism (the one with the most proteins in Uniprot). This resulted in our final protein data set with 669 organisms and over 72.000 proteins.

2.2 Domain annotation

Domains are often detected looking at two characteristics, its compactness and its extent of isolation. Interproscan 5 is a software package to scan sequence entries against the Interpro database, this database consists of 14 different databases like Pfam, SUPERFAMILY and PROSITE. Interpro is a database where all the information of sequence entries from other databases is saved in one place [18]. Pfam is a protein families database that uses hidden Markov models to iteratively detect families and domains in given sequences [19]. Hidden Markov models use known parameters to determine unknown parameters. SUPERFAMILY also is a

database that uses hidden Markov models to annotate proteins and genomes [20]. PROSITE is a manually curated database which consists of protein families and domains that reliably identifies these families and domains for given sequences [21].
Interpro is used for a range of different researches, such as checking if certain domains are present in proteins [22] [23], the further definition of genes [24] and confirmation of identified genes [25].
In this research, two scans were done using Interproscan 5. The first scan was just using the Pfam database and the second scan was using all the different databases provided by Interproscan 5. To obtain the results, the Pfam scan was used to be able to compare results found with the results of Schaefer *et al*, who also used the Pfam database. The second scan was done for possible future research, which could possibly lead to novel insights.

2.3 Storage

To store all information retrieved by the Interproscan service, we made use of the already existing polyQ2.0 database relational structure [14]. The polyQ2.0 database is an improved mySQL database for human polyglutamine proteins with various protein annotations, like gene information, Pfam domains and 3D structures. For this research, we only need a part of this information, so we used its Pfam subdatabase, and expanded it with a few columns: Organism, proteinId, Swiss-Prot/TrEMBL (wether the protein is part of Swiss-Prot or TrEMBL) and Interpro accession.

2.4 Domain analysis

Analyzing the correlation of domains found in polyQ proteins and polyQ protein frequency, we used Spearman's rank correlation coefficient. Spearman's rank correlation is used to measure the rank correlation of two lists of variables. In contrast to the Pearson correlation, which uses the values of the variables in the list to calculate the correlation, Spearman's rank correlation does not use the values but the ranks of the variables.
There are two formulas to calculate this correlation, one that can be used on all sets of variables, and one, more popular, that can only be used when all ranks are distinct integers, meaning no two variables in the same list can be the same.
For this research, following the methods of the 2012 paper, the following formula was used to calculate the Spearman correlation

$$1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

where d_i is the difference between the two ranks of variable i, and n is the number of observations.
The outcome of this formula is denoted as the Greek letter ρ (rho) and can be anything between -1 and +1. When ρ is equal to -1 or +1, the two variables have respectively negative or positive correlation. The 2012 paper states that any ρ above 0.3 is an indication that these domains are significantly enriched in polyQ proteins, this will also be used further on in this paper.

2.5 Gene ontology analysis

To analyze the significant domains found using the Spearman's rank correlation, we used gene ontology annotations [26] [27]. Gene ontology (GO) is an ontology made to annotate genes and gene products for all species in one place. This ontology can be divided into three domains of gene functions: Molecular Function, Biological Process and Cellular Component. Each entry in GO is connected to one of these three domains, where molecular function stands for the function on a molecular level, such as DNA-binding or catalysis, biological process stands for a specific goal that is genetically programmed to achieve, and a cellular component is a part of a cell. This ontology can be used with genes, proteins and domains. This makes it useful for this research, as both proteins and domains are being analyzed.
For the enrichment analysis, gProfiler was used. gProfiler is a web tool that calculates statistical enrichment for any given list of genes or proteins [28].
For the domain analyzation, a different method was used, our list of domains consisted of Pfam ID's, which are not included in gProfiler. Pfam2go bridged this problem, a list of Pfam ID's with corresponding GO terms [29].

Chapter 3

Results

Diseases related to elongated polyglutamine tracts have been studied a lot, however, polyQ proteins themselves have not. Schaefer *et al* suggested a shift of focus to wild-type polyQ proteins to possibly find functions of these proteins. Since the research in 2012 an enormous amount of protein data has been sequenced. To see if experiments show different results or give novel ideas, we recreated part of the experiments done by Schaefer *et al* [11]. The table below shows the differences in the dataset from 2012 and the new dataset.

Table 3.1: Quantitative comparison between 2012 and this research
*No available data

	2012 paper	New results
Species	11	669
Proteins	*	151.500
Human proteins	86	396
Domains	*	3984
Significant domains	40	170

3.1 Fraction of polyQ proteins in different organisms

For this calculation, we use the same definition for polyQ proteins as before, with at least 10 consecutive Qs and at least eight polyQ proteins. We also use the threshold of at least 750 proteins in the Swissprot/TrEMBL database and a fully sequenced genome as criteria for organisms.
Firstly, we will compare the fractions of organisms found by Schaefer *et al* and fractions for the same organisms we found in Figure 3.1. At first we observe a fairly similar figure to the figure of Schaefer *et al*, but upon further inspection, most fractions have become smaller, e.g. in *Homo sapiens* (0.34% to 0.29%), *Dictyostelium discoideum* (10.5% to 7.8%) and *Drosophila melanogaster* (3.8% to 2.6%). This reduction could have multiple explanations, one of which is our inclusion of TrEMLB proteins in our experiments, which increases the total amount of proteins per organism. Another reason could be the increase in proteins found due to alternative splicing [30], which also increases the total amount of proteins per organism. To compare the fraction of polyQ proteins with Schaefer *et al*, we created a figure with the same organisms as Figure 2 of [11]. Some organisms

are left out because these were not in our protein data set, *Arabidopsis thaliana*, *Ashbya gossypii*, *Bacillus subtilis*, *Archaeoglobus fulgidus*, *Methanocaldococcus jannaschii* and *Acanthamoeba polyphaga mimivirus*. These organisms had less than eight polyQ proteins in our database, so they were excluded from the final database.

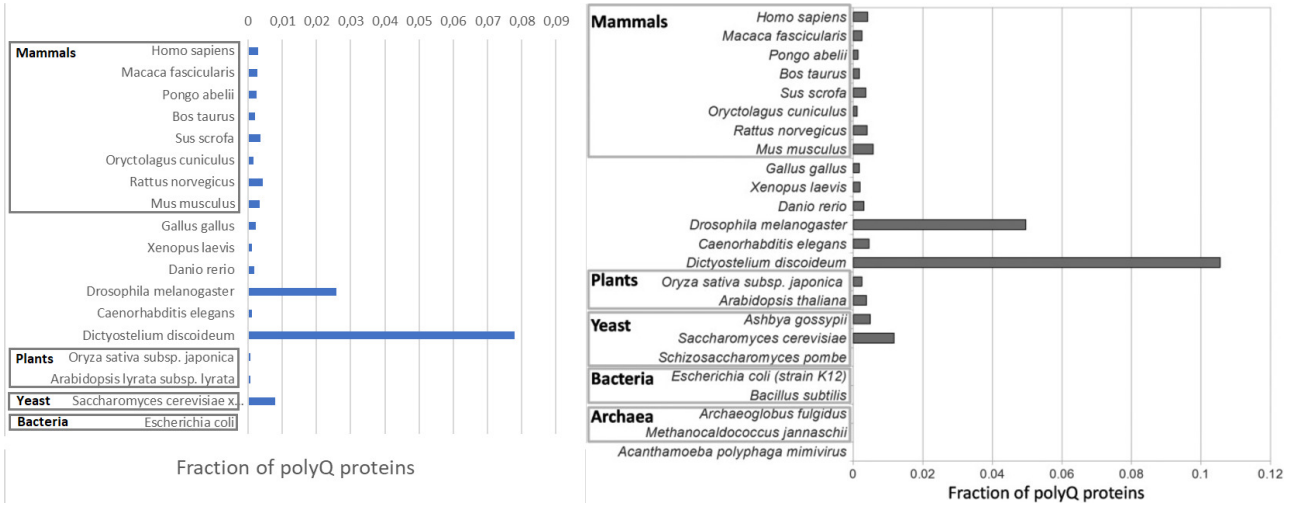

Figure 3.1: (left) Relative amount of polyQ proteins for several organisms in the data set, chosen to be comparable to the similar figure in the 2012 paper of Schaefer *et al* (right) [11]. A figure with all organisms can be found in the supplementary data

3.2 Domain correlation & function

To find out functions of polyq proteins, we studied the presence of domains in eukaryotic species to find over represented domains. Finding overrepresented domains is a good indication of the function of proteins due to the conserved function of domains over proteins and species. We calculated correlations of 3545 domains over eukaryotic species with over 5000 protein entries in the Pfam database using the Spearman correlation. We found 170 domains with a ρ over 0.3, which is more than four times the amount found by Schaefer *et al*, [11].

Table 3.2: Correlation of domains over eukaryotic species using the Spearman's rank correlation

Pfam ID	Pfam domain	Spearman's rank correlation
PF00069	Protein kinase domain	0.699
PF00076	RNA recognition motif	0.566
PF00271	Helicase conserved C-terminal domain	0.549
PF00176	SNF2 family N-terminal domain	0.544
PF00096	Zinc finger, C2H2 type	0.535
PF00439	**Bromodomain**	**0.496**
PF00169	**PH domain**	**0.49**
PF00018	SH3 domain	0.48
PF00505	HMG (high mobility group) box	0.479
PF00651	BTB/POZ domain	0.476
PF00620	**RhoGAP domain**	**0.469**
PF00621	RhoGEF domain	0.467
PF12796	Ankyrin repeats (3 copies)	0.464
PF00443	Ubiquitin carboxyl-terminal hydrolase	0.461
PF18016	SAM domain (Sterile alpha motif)	0.46
PF00320	GATA zinc finger	0.456
PF00010	Helix-loop-helix DNA-binding domain	0.451
PF00412	LIM domain	0.447
PF01388	ARID/BRIGHT DNA binding domain	0.447
PF13637	Ankyrin repeats (many copies)	0.446

List of the top 20 (out of 170) Pfam domains with a ρ of 0.3 or higher. Highlighted domains are also found in the 2012 results, in total 12 protein domains were found significant in both results.

Just over two-thirds of these domains have a molecular function in molecule binding, where the largest groups are DNA binding and ion binding. In all DNA binding domains, nine domains bind DNA, nine more bind DNA within a regulatory region to regulate transcription and six more domains that are sequence-specific, four more domains are RNA binding. Of the remaining domains, most are catalysts for several different reaction, like hydrolysis, phosphate groups and a few other reactions. Looking at the Biological processes of these domains, we can see which processes are driven by the molecular functions of the domains. Of course, lots of domains regulate transcription through DNA/RNA binding, but there are also processes that transduce signals, (de)phosphorylate proteins, deubiquitination and transmembrane transportation.

The results of Schaefer *et al* suggest the significant domains have roles in phosphatidylinositol signalling, protein degradation and molecular interactions [11]. However, to compare the results found with the results of Schaefer *et al*, we firstly analysed both our Pfam domains and the domains found by Schaefer *et al* using Pfam2GO, as in the 2012 paper did not attempt to classify results in a systematic way, listing only 14 of the 40 found domains and their functions. Analysing both results with the same tool, the comparison of these results is more precise than analysing the results gained from different tools. The tool we used is Pfam2GO, which converts the Pfam ID's to Gene ontology ID's, this tool gives two lists of GO terms we compare. Most of the functions found by Schaefer *et al* are also found in the new results, such as protein binding, zinc, nucleic acid and ion binding and signal transduction. One big difference is again the amount of GO ID's found, 23 in 2012 versus 160 now, this is expected because more Pfam domains have been found compared to 2012.

Schaefer found three domains with functions in the phosphatidylinositol signaling system, which do not show up in the new results, this suggests these domain functions are not as important as was mentioned in the 2012

paper. However, the biggest difference between both results sets is the size, having found many more functions of domains overrepresented in polyQ proteins. These novel findings show that these new domain functions mostly are DNA or RNA binding, suggesting polyQ proteins seem to have, besides the findings of Schaefer *et al*, important roles in transcription modulation.

3.3 Protein function

There already has been research showing polyQ proteins have functions related to transcriptional regulation [11]. The newly found results appear to confirm this. The database showed 364 human polyQ proteins, but after further analysis we found 61 genes alternatively spliced to get 364 proteins. This is roughly the same amount of human proteins found as in 2012. Most of these proteins have functions related to DNA and RNA binding, and some bind cyclic compounds and chromatin. There are some differences in results found, but the general functions found have not changed, it appears polyQ proteins mostly function in the regulation of transcription through DNA binding.

Table 3.3: Significant Gene Ontology ID's showing the molecular functions(MF) of the human polyQ proteins, the tables with the cellular components(CC) and biological processes(BP) are added in the supplementary data

Gene Ontology ID	**Gene Ontology name**
GO:0140110	transcription regulator activity
GO:0044212	transcription regulatory region DNA binding
GO:0001067	regulatory region nucleic acid binding
GO:0003677	DNA binding
GO:0003700	DNA-binding transcription factor activity
GO:0000981	DNA-binding transcription factor activity, RNA polymerase II-specific
GO:0001012	RNA polymerase II regulatory region DNA binding
GO:0000976	transcription regulatory region sequence-specific DNA binding
GO:1990837	sequence-specific double-stranded DNA binding
GO:0003676	nucleic acid binding
GO:0000977	RNA polymerase II regulatory region sequence-specific DNA binding
GO:0003690	double-stranded DNA binding
GO:0003682	chromatin binding
GO:0000987	proximal promoter sequence-specific DNA binding
GO:0043565	sequence-specific DNA binding
GO:0003712	transcription coregulator activity
GO:0003713	transcription coactivator activity
GO:0000978	RNA polymerase II proximal promoter sequence-specific DNA binding
GO:0008134	transcription factor binding
GO:1901363	heterocyclic compound binding
GO:0097159	organic cyclic compound binding
GO:0140297	DNA-binding transcription factor binding
GO:0001228	DNA-binding transcription activator activity, RNA polymerase II-specific
GO:0001216	DNA-binding transcription activator activity
GO:0061629	RNA polymerase II-specific DNA-binding transcription factor binding
GO:0001085	RNA polymerase II transcription factor binding
GO:0043425	bHLH transcription factor binding

To see if any connections can be made in molecular function of the human proteins that cause neurodegenerative diseases when the glutamine tracts are elongated, a list of disease causing proteins was made using the 2014

paper of Fan *et al.* [31]. This resulted in a list containing the nine known polyQ diseases and their genes. Surprisingly, these proteins do not seem to play a role in the DNA binding like the majority of polyQ proteins. Only SCA17 and SBMA interact with a general transcription initiation factor. The other pathogenic proteins are cellular components, specifically in the nucleus. To see if the non pathogenic proteins have the same functions as the pathogenic proteins, one more test was done using gProfiler, this time with only the non pathogenic proteins. The results of this test were compared with the results of first test, showing only one shared function in both groups, being the cellular component nucleoplasm.

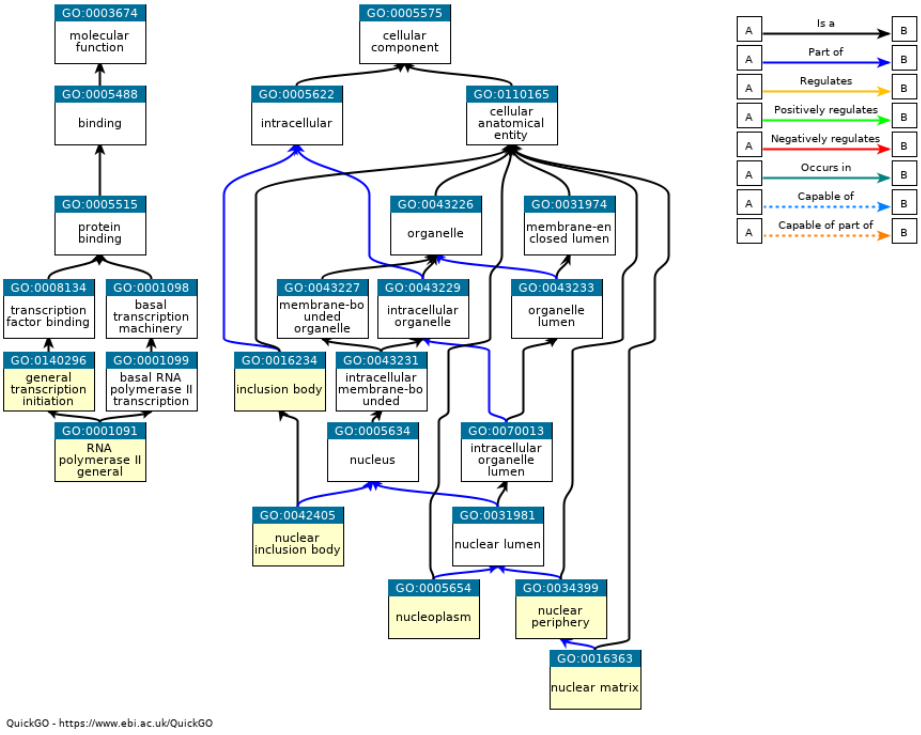

Figure 3.2: The ancestor chart of the GO terms of the nine neurodegenerative disease proteins using QuickGO [32]. The boxes marked yellow are the GO terms of the disease proteins.

Because there are only nine disease causing proteins, the sample size is too small to get a proper enrichment analysis using gProfiler. For this reason, we also looked at the Gene Ontology terms found in these proteins, this resulted in 115 unique GO terms, 11 of which occur more than once, shown in table 3.3. As seen in this table, only two of the GO terms are also found significantly enriched using gProfiler. Although these two results differ from each other, both of these results mostly show GO terms with functions as cellular components.

Table 3.4: GO terms occurring more than once in the nine disease causing proteins, highlighted domains are also found in the gProfiler results

Gene Ontology ID	**Gene Ontology name**	**number of occurrences**
GO:0003677	DNA binding	4
GO:0005623	cell	2
GO:0005634	nucleus	4
GO:0005654	**nucleuplasm**	**4**
GO:0005730	nucleolus	2
GO:0005737	cytoplasm	3
GO:0005829	cytosol	2
GO:0016363	**nuclear matrix**	**2**
GO:0030425	dendrite	2
GO:0042802	identical protein binding	2
GO:0048471	perinuclear region of cytoplasm	2

Chapter 4

Conclusions

In conclusion, as seen in the results, the fraction of polyQ proteins has not changed relatively to the fractions found in the 2012 article, except the amount of proteins found has increased a lot, so it was interesting to see if there were any changes in protein or domain function, since many new polyQ proteins have been discovered. Starting with the Spearmans rank correlation, 170 domains were found with a rho of over 0.3 making them significant, compared to 40 domains found in 2012, this on itself already suggests polyQ proteins have more functions than initially thought. To verify this, we looked at the functions of these domains using GO, most of the domains had functions in molecule binding, such as DNA and ion binding. Comparing these results with Schaefers results, we find almost all of the functions found by Schaefer, and many more, most of these new functions found are DNA binding. In the 2012 paper, there is a focus on the phosphatidylinositol signaling system, remarkably, this function is absent in our results, which suggests that polyQ proteins actually do not significantly play a role in the phosphatidylinositol signaling system. The other results from 2012 absent in the new results are: RAN GTPase binding, vesicle-mediated transport, and phosphatase activity. The protein functions in humans are in line with the domain functions, being mostly functional in DNA and RNA binding. This is likely the case because of our 364 human proteins, only 61 genes were found, this is roughly the same amount found in 2012, so it seems not many new human polyQ proteins have been found resulting in approximately the same results. The functions of disease causing proteins have also been investigated separately to see if any connections could be made between protein function and disease cause. The nine proteins did not seem to play a role in protein binding, only two proteins had enriched GO(Molecular Function) terms in transcription initiation factor binding, SCA17 and SBMA. The other proteins showed no enriched GO(MF) terms, probably due to the small sample size as mentioned before. More detailed research of these proteins is necessary to find the difference between disease causing polyQ proteins and non-disease causing polyQ proteins.

4.1 Further research

These new findings show that our understanding of polyQ proteins have changed over the years, and is still very relevant to research. In this research, the focus was mainly on the domain and protein function to see if the results had changed since 2012. Seeing as the results have changed, the next step could be revisiting the other results shown in the 2012 paper, like the functions of proteins interacting with polyQ proteins and polyQ emergence in protein families. Also, with Interproscan 5, we created a bigger database with all tools available, further research could repeat this research but with all data found.

Bibliography

[1] M. Buljan and A. Bateman, "The evolution of protein domain families," 2009.

[2] J. H. Laity, B. M. Lee, and P. E. Wright, "Zinc finger proteins: new insights into structural and functional diversity," *Current opinion in structural biology*, vol. 11, no. 1, pp. 39–46, 2001.

[3] "The cause of huntington's disease," `http://www.ehdn.org/about-hd/` [Online; accessed 12-12-2019].

[4] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. Tosatto, and R. D. Finn, "The Pfam protein families database in 2019," *Nucleic Acids Research*, vol. 47, pp. D427–D432, 10 2018.

[5] R. E. Tanzi, A. I. McClatchey, E. D. Lamperti, L. Villa-Komaroff, J. F. Gusella, and R. L. Neve, "Protease inhibitor domain encoded by an amyloid protein precursor mrna associated with alzheimer's disease," *Nature*, vol. 331, no. 6156, p. 528, 1988.

[6] K. Pawlowski, F. Pio, Z.-L. Chu, J. C. Reed, and A. Godzik, "Paad–a new protein domain associated with apoptosis, cancer and autoimmune diseases," *Trends in biochemical sciences*, vol. 26, no. 2, pp. 85–87, 2001.

[7] R. Bonasio, E. Lecona, and D. Reinberg, "Mbt domain proteins in development and disease," in *Seminars in cell & developmental biology*, vol. 21, pp. 221–230, Elsevier, 2010.

[8] F. O. Walker, "Huntington's disease," *The Lancet*, vol. 369, no. 9557, pp. 218–228, 2007.

[9] C. A. Ross, "Intranuclear neuronal inclusions: a common pathogenic mechanism for glutamine-repeat neurodegenerative diseases?," *Neuron*, vol. 19, no. 6, pp. 1147–1150, 1997.

[10] D. C. Rubinsztein and J. Carmichael, "Huntington's disease: molecular basis of neurodegeneration," *Expert Reviews in Molecular Medicine*, vol. 5, no. 20, p. 121, 2003.

[11] M. H. Schaefer, E. E. Wanker, and M. A. Andrade-Navarro, "Evolution and function of cag/polyglutamine repeats in protein–protein interaction networks," *Nucleic acids research*, vol. 40, no. 10, pp. 4273–4287, 2012.

[12] J. Hall, K. Heel, and R. McCauley, "Glutamine," *British Journal of Surgery*, vol. 83, no. 3, pp. 305–312, 1996.

[13] M. Nance, V. Mathias-Hagen, G. Breningstall, M. Wick, and R. McGlennen, "Analysis of a very large trinucleotide repeat in a patient with juvenile huntington's disease," *Neurology*, vol. 52, no. 2, pp. 392–392, 1999.

[14] C. Li, J. Nagel, S. Androulakis, J. Song, and A. M. Buckle, "Polyq 2.0: an improved version of polyq, a database of human polyglutamine proteins," *Database*, vol. 2016, 2016.

[15] K. Forslund and E. L. Sonnhammer, "Predicting protein function from domain content," *Bioinformatics*, vol. 24, no. 15, pp. 1681–1687, 2008.

[16] U. Consortium, "Uniprot: a worldwide hub of protein knowledge," *Nucleic acids research*, vol. 47, no. D1, pp. D506–D515, 2018.

[17] E. W. Sayers, R. Agarwala, E. E. Bolton, J. R. Brister, K. Canese, K. Clark, R. Connor, N. Fiorini, K. Funk, T. Hefferon, *et al.*, "Database resources of the national center for biotechnology information," *Nucleic acids research*, vol. 47, no. Database issue, p. D23, 2019.

[18] A. L. Mitchell, T. K. Attwood, P. C. Babbitt, M. Blum, P. Bork, A. Bridge, S. D. Brown, H.-Y. Chang, S. El-Gebali, M. I. Fraser, *et al.*, "Interpro in 2019: improving coverage, classification and access to protein sequence annotations," *Nucleic acids research*, vol. 47, no. D1, pp. D351–D360, 2018.

[19] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, *et al.*, "The pfam protein families database in 2019," *Nucleic acids research*, vol. 47, no. D1, pp. D427–D432, 2018.

[20] J. Gough, K. Karplus, R. Hughey, and C. Chothia, "Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure," *Journal of molecular biology*, vol. 313, no. 4, pp. 903–919, 2001.

[21] C. J. Sigrist, E. De Castro, L. Cerutti, B. A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, and I. Xenarios, "New and continuing developments at prosite," *Nucleic acids research*, vol. 41, no. D1, pp. D344–D347, 2012.

[22] M. T. Mata, A. Palma, C. García-Gómez, M. López-Parages, V. Vázquez, I. Cheng-Sánchez, F. Sarabia, F. López-Figueroa, C. Jiménez, and M. Segovia, "Type ii-metacaspases are involved in cell stress but not in cell death in the unicellular green alga dunaliella tertiolecta," *Microbial Cell*, vol. 6, no. 11, p. 494, 2019.

[23] P. Jiang, J. Shao, and L. G. Nemchinov, "Identification of emerging viral genomes in transcriptomic datasets of alfalfa (medicago sativa l.)," *Virology Journal*, vol. 16, no. 1, pp. 1–12, 2019.

[24] Z. Zhang, W. Liu, Z. Ma, W. Zhu, and L. Jia, "Transcriptional characterization and response to defense elicitors of mevalonate pathway genes in cotton (gossypium arboreum l.)," *PeerJ*, vol. 7, p. e8123, 2019.

[25] E. S. Bartholomew, K. Black, Z. Feng, W. Liu, N. Shan, X. Zhang, L. Wu, L. Bailey, N. Zhu, C. Qi, *et al.*, "Comprehensive analysis of the chitinase gene family in cucumber (cucumis sativus l.): From gene identification and evolution to expression in response to fusarium oxysporum," *International journal of molecular sciences*, vol. 20, no. 21, p. 5309, 2019.

[26] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, p. 25, 2000.

[27] G. O. Consortium, "The gene ontology resource: 20 years and still going strong," *Nucleic acids research*, vol. 47, no. D1, pp. D330–D338, 2018.

[28] U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, and J. Vilo, "g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)," *Nucleic acids research*, 2019.

[29] M. et al., "pfam2go: Add go terms based on pfam accessions," *Nucleic acids research*, 2015.

[30] B. Modrek and C. Lee, "A genomic view of alternative splicing," *Nature genetics*, vol. 30, no. 1, p. 13, 2002.

[31] H.-C. Fan, L.-I. Ho, C.-S. Chi, S.-J. Chen, G.-S. Peng, T.-M. Chan, S.-Z. Lin, and H.-J. Harn, "Polyglutamine (polyq) diseases: genetics to treatments," *Cell transplantation*, vol. 23, no. 4-5, pp. 441–458, 2014.

[32] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'donovan, and R. Apweiler, "Quickgo: a web-based tool for gene ontology searching," *Bioinformatics*, vol. 25, no. 22, pp. 3045–3046, 2009.