Between Interactions and Aggregates: The PolyQ Balance

Pablo Mier ,* and Miguel A. Andrade-Navarro

Faculty of Biology, Institute of Organismic and Molecular Evolution, Johannes Gutenberg University of Mainz, Mainz, Germany

*Corresponding author: E-mail: munoz@uni-mainz.de.

Accepted: 27 October 2021

Abstract

Polyglutamine (polyQ) regions are highly abundant consecutive runs of glutamine residues. They have been generally studied in relation to the so-called polyQ-associated diseases, characterized by protein aggregation caused by the expansion of the polyQ tract via a CAG-slippage mechanism. However, more than 4,800 human proteins contain a polyQ, and only nine of these regions are known to be associated with disease. Computational sequence studies and experimental structure determinations are completing a more interesting picture in which polyQ emerge as a motif for modulation of protein–protein interactions. But long polyQ regions may lead to an excess of interactions, and produce aggregates. Within this mechanistic perspective of polyQ function and malfunction, we discuss polyQ definition and properties such as variable codon usage, sequence and context structure imposition, functional relevance, evolutionary patterns in species-centered analyses, and open resources.

Key words: homorepeat, polyglutamine, CAG-expansion diseases, aggregation, protein–protein interaction.

Significance

Usually, polyglutamine (polyQ) regions are studied in relation to human neurodegenerative diseases. However, there are more than 4,800 human proteins with these regions, and only nine of them are related to a disease, which makes them important beyond the scope of the diseases they are associated with. Here, we summarize the known information about polyQ as sequence motifs, in perspective of a delicate balance between the role of polyQ in modulating protein–protein interactions and polyQ pathological aggregation.

Introduction

Amino acid homorepeats are stretches of consecutive identical amino acids in protein sequences. They are abundant in eukaryotic proteomes, and can be formed by almost any amino acid (Faux et al. 2005; Chavali et al. 2020). They have a general low profile in terms of sequence motif research, due to their apparent simplicity and historical lack of known/expected function (Karlin and Burge 1996; Huntley and Golding 2000; Lavorgna et al. 2001).

When formed by glutamine residues, these homorepeats are called polyglutamine or polyQ regions. PolyQ regions are the most frequent homorepeats in eukaryotic species (Lobanov and Galzitskaya 2012). They were first described in 1985 as "unusual structures", as part of the wheat alpha/beta-gliadin proteins (Sumner-Smith et al. 1985) and the opa repetitive element in the *Drosophila* Notch gene (Wharton et al. 1985). They were not characterized in any human protein until 1989, in the N-terminal domain of the human androgen receptor (Faber et al. 1989).

The first report of their association with a human disease was in 1992, when the expansion of the polyQ in the first exon of the human androgen receptor gene was correlated with the Spinal and Bulbar Muscular Atrophy disease (La Spada et al. 1992). Since then, eight more human genetic diseases have been described to be associated with the expansion of a polyQ region (see, e.g., Gatchel and Zoghbi 2005), often leading to nuclear protein aggregates containing the expanded protein or fragments in neurons (Ross 1997). As a result, research on polyQ regions has been biased toward the relation of the polyQ-containing protein to human disease. As a reference, as of October 2021, 85% of publications indexed in PubMed with the terms "polyglutamine" or "polyQ" were also related to "disease". This is highly unbalanced considering that there are 4,808 human proteins with a polyQ region (see below) and in only nine of them (0.2%) the polyQ expansion is associated with disease.

Here we summarize current knowledge about polyQ regions, covering from the most basic nucleotide level to their protein structure and function.

Definition

PolyQ regions are defined as protein regions with a very high proportion of glutamine residues. However, there is currently no consensus in the literature about what this proportion should be. The detection of polyQ regions in proteins is based on thresholds (≥*X*/*Y*; where *X* is the minimum number of glutamine residues to be found in a local region of a given length *Y*). Up to nine different thresholds have been used in different studies or databases to define and detect polyQ regions: ≥4/4 (Mier et al. 2020), ≥5/5 (Albà and Guigó 2004), ≥6/6 (Lobanov and Galzitskaya 2012), ≥7/7 (Jorda and Kajava 2010), ≥4/5 (The UniProt Consortium 2019), ≥4/6 (Mier and Andrade-Navarro 2017), ≥4/7 (Faux et al. 2005), ≥8/10 (Mier et al. 2017), and ≥9/10 (Schaefer et al. 2012).

The election of a threshold is crucial because it will impact downstream analysis, as the polyQ regions detected will differ. Several thresholds share the minimum of four glutamines, because to have a structurally functional polyQ it was described that at least four glutamines are necessary (Totzeck et al. 2017). This work is based on structural analysis of insertion points of polyQ according to comparisons between proteins with known structure and their orthologs with polyQ. The equivalent positions of insertion of polyQ regions of four or more glutamines in orthologs without polyQ display the characteristic helical/random coil structure context that is connected to polyQ functionality, as we will describe later.

To have a comprehensive overview of the length distribution and purity of the polyQ regions in the human proteome, we used all of the thresholds above to look for polyQ regions in the complete version of the human proteome in the UniProtKB database (release 2020_06). We used a standalone version of the sQanner tool (Mier et al. 2020). It is clear that results are mainly influenced by the purity of the polyQ (table 1). Detailed lists of proteins found with each of the polyQ lengths and thresholds (including the position of the polyQ in the sequence) can be found in supplementary file S1, Supplementary Material online.

Variable Codon Usage

Glutamine residues are coded by codons CAG and CAA. In *Homo sapiens*, CAG is enriched in an unbalanced 3:1 proportion over CAA (Komar 2016). Glutamines part of polyQ regions are more enriched in CAG than expected, in a length-dependent way: longer polyQ are more enriched in CAG than shorter regions (Galzitskaya et al. 2019). This is true for all mammals (Mier and Andrade-Navarro 2018). Furthermore, euteleostomes (bony vertebrates) share the CAG enrichment over CAA, whereas model organisms such as the tunicate *Ciona intestinalis*, the worm *Caenorhabditis elegans*, and the yeast *Saccharomyces cerevisiae* have it shifted to a 3:1 proportion CAA:CAG. Interestingly, the coding of polyQ regions in *Drosophila melanogaster*, whereas closer taxonomically to *C. elegans*, resemble that of the euteleostomes (Mier and Andrade-Navarro 2018).

PolyQ tracts in proteins associated with disease are even more enriched in CAG, reaching almost exclusive CAG regions (Nalavade et al. 2013). CAA insertions in a *Drosophila* model of a polyQ-associated disease show that, even though the produced polyQ is of the same length, these proteins are of reduced toxicity (Li et al. 2008). This suggests that there is a component of CAG-expanded RNA toxicity, at least in this experimental setup. However, although there is evidence that the CAG/CAA composition of polyQ-encoding transcripts is related to disease (Genetic Modifiers of Huntington's Disease Consortium 2019; Wright et al., 2019), there is at least one disease caused by CAG expansion in a nontranslated region (SCA12; Holmes et al. 1999; Lone et al. 2016; Srivastava et al. 2017). This highlights the importance of CAG-expanded RNA toxicity in humans (e.g., Bhambri et al. 2020).

PolyQ regions are sometimes composed by glutamine plus other inserted residues (impurities) (Mier et al. 2020). They are usually proline, histidine, and leucine residues (Ramazzotti et al. 2012). The first two are specially enriched as polyQ impurities probably because they are coded by codons one mutational event apart from CAG or CAA (proline enriched in

Table 1
Number of PolyQ Regions in the Human Proteome (UniProt release 2020_06)

Pure polyQ	
Length	**Number**
4	744
5	260
6	134
7	73
8	57
9	19
10	49
>11	165
Impure polyQ	
Threshold	**Number**
≥4/5	2,950
≥4/6	4,808
≥4/7	6,871
≥8/10	397
≥9/10	291

codons CCG and CCA, histidine in CAC and CAT) (Ramazzotti et al. 2012). Although leucine may be coded by CTA and CTG, both also one mutation away from CAA and CAG, the most usual codons for leucine when present as impurity in polyQ are CTC/CTT, and therefore similarity to CAG/CAA codons does not explain the presence of leucine within the polyQ regions.

Sequence and Context Structure Imposition

It was found that isolated polyQ peptides with variable lengths are not particularly different in their structural properties and aggregation propensity (Klein et al. 2007). This is in apparent contradiction to the fact that expanded polyQ causes disease. The solution to this contradiction is that the sequence context of the polyQ plays a role in its properties.

PolyQ regions are generally preceded by helical conformation and followed by a random coil (Totzeck et al. 2017). A recent NMR structural study of various polyQ variants with their sequence context in Huntingtin demonstrated that the polyQ prolongs the preceding helical conformation (Urbanek et al. 2020); it was shown previously that the helical structure of polyQ is stabilized by intrahelical hydrogen bonds mediated by the side chains of the glutamine residues (Zhemkov et al. 2016; Escobedo et al., 2019). In Huntingtin and other proteins, polyQ is followed C-terminally by a polyP or Pro-rich region. It was observed using synthetic peptides that a polyP C-terminal from a polyQ decreases its propensity to aggregate, whereas it does not produce such an effect if situated N-terminally (Bhattacharyya et al. 2006). Evolutionary studies of polyQ context show that indeed such a dependency can be observed in natural proteins: considering protein families, polyQ may occur at a particular position, and while proteins in the family with the polyQ may display a polyP following, the polyP alone is not observed (Mier et al. 2017).

Examination of the immediate vicinity of polyQ regions revealed the prevalence of leucine at positions -1 and -5 from the polyQ and of prolines right after (fig. 1*a*) (Ramazzotti et al. 2012; Mier et al. 2020). These observations support the propensity for the preceding alpha-helical structure to be transmitted to the polyQ, and the known polyP bias following it, respectively. Both compositional biases reflect the existence of evolutionary pressure that results in conformational restrictions against polyQ aggregation-induced toxicity (Bhattacharyya et al. 2006; Darnell et al. 2007; Eftekharzadeh et al. 2016; Shen et al. 2016; Urbanek et al. 2020).

Biophysical Studies of PolyQ Structure

The structure of polyQ and their aggregates have been investigated in multiple biophysical studies. PolyQ regions display very different structural behavior depending on their sequence context (Whi Kim et al., 2009; Chavali et al. 2020), thus it is difficult to compare studies that focus on synthetic polyQ, fragments of proteins containing polyQ, or entire polyQ-containing proteins. PolyQ aggregates were proposed to be composed of multimers of anti-parallel beta-strands (Perutz et al. 1994) or beta turns (Buchanan et al. 2014). Regarding polyQ regions in monomeric state, studies reported that they adopt collapsed structures in an ensemble of globular structures (Crick et al. 2006), or random coil structure (Bennett et al. 2002; Masino et al. 2002).

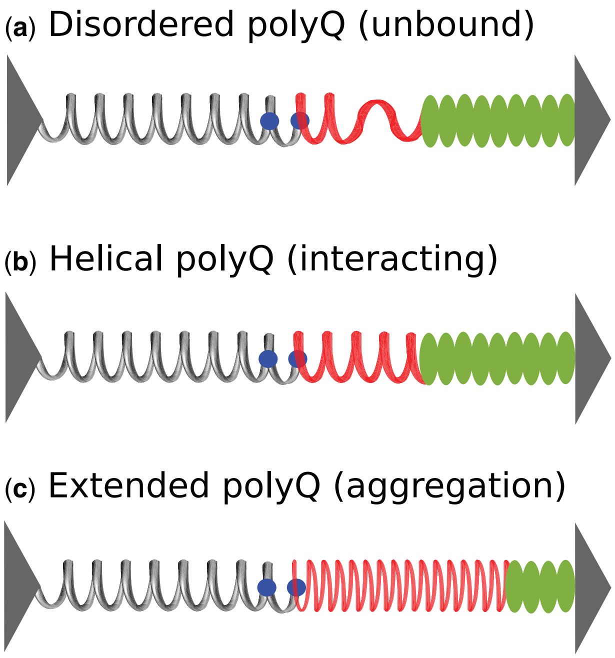
(a) Disordered polyQ (unbound)
(b) Helical polyQ (interacting)
(c) Extended polyQ (aggregation)

Fig. 1.—PolyQ context and structure. PolyQ regions (red) are usually positioned C-terminally to a coiled coil (gray) and N-terminally to a polyP or Proline-rich region (green). Leucine residues (blue) are found typically at positions -1 and -5 from the polyQ. PolyQ regions can be either (*a*) disordered, (*b*) in helical conformation to aid protein–protein interaction, or (*c*) forming beta-aggregates.

The differences between the studies mentioned reflect the importance of sequence context for the structure and aggregation state of polyQ regions, as demonstrated using dynamics simulations (Ruff et al. 2014), or solubilizing polyQ regions by fusing them to protein sequences (Scherzinger et al. 1997). A recent study demonstrates how the insertion of just a few glutamines in glutamine-rich beta-hairpin monomers destroy their secondary structure (Siu et al. 2021). Together these studies demonstrate the complexity of structures adopted by polyQ, with properties very far from globular proteins.

Functional Relevance of PolyQ Regions

It has been observed that polyQ is often inserted in evolution after a coiled-coil region (Schaefer et al. 2012) or regions with helical structure, and preceding a disordered region (Totzeck et al. 2017). Coiled coils are motifs used for protein–protein interaction. Together with the propensity of long polyQ to form aggregates, this suggested a role for polyQ in the modulation of protein–protein interactions. This was further supported by the enrichment of polyQ in proteins with many interactions, and by the enrichment of coiled coils in proteins that interact with polyQ proteins (Schaefer et al. 2012). It has been hypothesized that upon interaction, the flexible polyQ would adopt a more ordered helical conformation extending the preceding helix and thus making the interaction stronger (fig. 1*b*). Although the expansion of CAG repeats facilitates the evolutionary emergence of polyQ regions, the same mechanism can lead to genetic disease when the excessively expanded polyQ affects negatively the modulated interaction, resulting in aggregates (fig. 1*c*) (Gatchel and Zoghbi 2005; Kuiper et al. 2017) and dysfunctional interactions (Hosp et al. 2017; Zhao et al. 2018). This is avoided by disrupting the repeat by mutation to CAA, probably fixing the polyQ to an optimal length (Albà et al. 2001; Mier and Andrade-Navarro 2020).

The enrichment of polyQ regions at the N-terminal of regions adopting helical and coiled-coil structure suggests that their emergence in certain protein families is functionally related to them (Fiumara et al. 2010; Schaefer et al. 2012). Moreover, the aggregates of a protein with a pathogenic polyQ increase when interacting with coiled-coil containing partners but decreases when interacting with partners not predicted to contain coiled coils. This finding suggested that the formation of aggregates is related to an alteration of the normal function of polyQ in coiled-coil mediated protein interactions (Petrakis et al. 2013). The enrichment of polyQ regions in proteins with many interaction partners (Pelassa and Fiumara 2015), and the observation of a significant number of families with multiple independent events of emergence of polyQ (Schaefer et al. 2012), suggested that particular protein families have a selective advantage in having polyQ. Collectively, with the studies that have demonstrated experimentally the dependency of coiled-coil protein interactions with polyQ regions (e.g., Petrakis et al. 2012; Ashkenazi et al. 2017; Kwon et al. 2018; see next section), these results suggest that polyQ is a motif that modulates protein interactions of helical motifs, and explain the formation of aggregates as a corruption of this function due to an abnormally long polyQ.

The evolutionary increase in pathway, organelle, and tissue complexity in multiple eukaryotic phyla is likely due to selective advantage resulting from organisms with increased regulatory features (Hedges et al. 2004). Thus, if the association of polyQ functionality to protein–protein interactions discussed above were its prevalent function, one could expect a certain increase of polyQ regions in relation to the increasing organismal complexity. It is true that polyQ have a relative low frequency in prokaryotic species compared with eukaryotes (Faux et al. 2005; Mier et al. 2017). However, polyQ is much more prevalent in unicellular species *Paramecium tetraurelia* and in the fungi *Kluyveromyces lactis* (69% and 27% of families with homorepeats have polyQ, respectively) than in human (6%) (Mier et al. 2017). For these reasons, polyQ can be expected to have other specialized roles.

Experimental Evidence of PolyQ Function in Protein Interactions

A number of studies have investigated the participation of polyQ regions in protein interactions. Coiled-coil-mediated interactions of Foxo with SCA3 and other polyQ proteins induced toxicity in the neurons of *D. melanogaster* (Kwon et al. 2018). The polyQ region of ataxin-3 participates in its interaction with beclin 1 (Ashkenazi et al. 2017). In another study, 21 interactors were found to alter the formation of aggregates induced by a toxic construct of ataxin-1 with an enlarged polyQ (Petrakis et al. 2012): a significant amount of enhancers of toxicity contained coiled coils, whereas none of the interactors reducing toxicity were found to contain coiled coils. In this study, it was found that the N-terminal coiled-coil domain of MED15 was enough to produce aggregation of ataxin-1 in vitro. Together, these studies demonstrate that polyQ regions are motifs that, similarly to intrinsically disordered domains, facilitate a great variety of interactions without the need of adopting a globular fixed structure. This role might be shared by other homorepeats such as polyA (Pelassa et al. 2014).

Multiple works have described the relation of polyQ regions with the activation of transcription (Gerber et al. 1994; Gemayel et al. 2015), in a length-dependent way (Atanesyan et al. 2012). This function, along with others (Lee et al. 2013; Chavali et al. 2017), is a consequence of their involvement in the modulation of protein–protein interactions.

Evolutionary Patterns in Species-Centered Analyses

PolyQ regions can emerge in a sequence either by point mutations (DNA substitution, deletion or insertion) or by DNA replication slippage (Albà et al. 2001). The latter is further associated with the previously mentioned polyQ growth via expansion of CAG triplets (Hancock 1996; Liu and Wilson 2012). In both cases they appear in a sequence in a structurally favorable environment; structures from orthologous sequences of human polyQ-containing proteins show that they emerge after a helical region (Totzeck et al. 2017). It seems plausible that glutamine residues appear randomly, start accumulating and are evolutionarily selected to extend and aid a preceding helix involved in protein–protein interaction.

In human proteins with CAG expansions leading to disease, the corresponding evolutionary expansion of polyQ can be observed when analyzing orthologs from species at

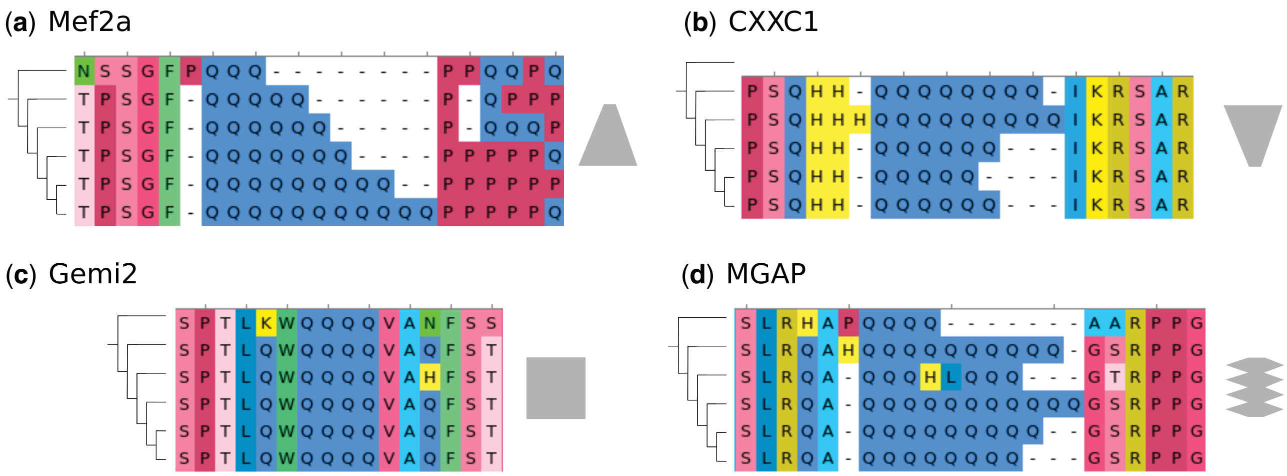
(a) Mef2a
N S S G F P Q Q Q - - - - - - - - P P Q Q P Q
T P S G F - Q Q Q Q Q - - - - - - P - Q P P P
T P S G F - Q Q Q Q Q Q - - - - - P - Q Q Q P
T P S G F - Q Q Q Q Q Q Q - - - - P P P P P Q
T P S G F - Q Q Q Q Q Q Q Q Q - - P P P P P P
T P S G F - Q Q Q Q Q Q Q Q Q Q Q P P P P P Q
(b) CXXC1
P S Q H H - Q Q Q Q Q Q Q Q - I K R S A R
P S Q H H H Q Q Q Q Q Q Q Q Q I K R S A R
P S Q H H - Q Q Q Q Q Q - - - I K R S A R
P S Q H H - Q Q Q Q Q - - - - I K R S A R
P S Q H H - Q Q Q Q Q Q - - - I K R S A R
(c) Gemi2
S P T L K W Q Q Q Q V A N F S S
S P T L Q W Q Q Q Q V A Q F S T
S P T L Q W Q Q Q Q V A H F S T
S P T L Q W Q Q Q Q V A Q F S T
S P T L Q W Q Q Q Q V A Q F S T
S P T L Q W Q Q Q Q V A Q F S T
(d) MGAP
S L R H A P Q Q Q Q - - - - - - - A A R P P G
S L R Q A H Q Q Q Q Q Q Q Q Q Q - G S R P P G
S L R Q A - Q Q Q H L Q Q Q - - - G T R P P G
S L R Q A - Q Q Q Q Q Q Q Q Q Q Q G S R P P G
S L R Q A - Q Q Q Q Q Q Q Q Q - - G S R P P G
S L R Q A - Q Q Q Q Q Q Q Q - - - G S R P P G

FIG. 2.—Examples of polyQ evolution patterns in orthologous proteins. The analysis is centered in the human species and considers a series of species at various taxonomic distances. Alignment of orthologs of proteins (*a*) Mef2a, (*b*) CXXC1, (*c*) Gemi2, and (*d*) MGAP from species *Gallus gallus* (top), *Bos taurus*, *Mus musculus*, *Macaca mulatta*, *Pongo abelii* and *H. sapiens* (bottom). The phylogenetic tree was built using taxonomic information from the NCBI Taxonomy resource Common Taxonomy Tree (Sayers et al. 2009) and the R library phyloseq version 1.29.0 (McMurdie and Holmes 2013). The complete alignments can be found in supplementary files S2–S5, Supplementary Material online.

Table 2

Resources Related to PolyQ Regions

Resource	Content	Reference
sQanner	Evaluation of the abundance of polyQ regions in a protein data set	Mier et al. (2020)
EvoPPI 1.0	Comparison of PPI data from several databases within and among ten species. The authors illustrate its use with a detailed case study using all nine polyQ-associated disease proteins	Vázquez et al. (2019)
HDNetDB	Interaction knowledge related to Huntington's disease, a polyQ-associated disease	Kalathur et al. (2017)
dAPE	Assessment of the evolution of homorepeats and their protein context	Mier and Andrade-Navarro (2017)
HRaDis	Database of associations between homorepeats and diseases	Lobanov et al. (2016)
PolyQ database	Information about 135 mouse models of polyQ-associated diseases, plus detailed descriptions about phenotypes and therapeutic approaches tested in vivo	Szlachcic et al. (2015)
HRaP	Database of occurrence of homorepeats and patterns in 122 proteomes (110 species)	Lobanov et al. (2014)

increasingly shorter distances to human. But, as discussed above, polyQ is present in many protein families that do not cause disease, and, for example, other primates have proteins with longer polyQ than the human ortholog (Mier and Andrade-Navarro 2018). We note that other patterns of polyQ evolution can be observed in a species-centric analysis, probably reflecting different selective pressure on particular proteins: increase, decrease, stable, and no pattern (fig. 2 and supplementary files S2–S5, Supplementary Material online).

A higher usage of the CAG codon in unstable polyQ in Amniota species (and not in Teleostei or Insecta) suggests that the CAG-slippage mechanism of polyQ expansion is specific to amniotes (Mier and Andrade-Navarro 2020). The general higher number of protein interactors in proteins with stable polyQ regions versus those with unstable polyQ, further supports the role of polyQ in protein interactions (Mier and Andrade-Navarro 2020).

Conclusions

PolyQ regions constitute an example of an overseen motif that may have been partially misrepresented in the literature, due to both its association to disease and difficulty to study. Increasing genomic information and experimental evidence, significantly from high-throughput interaction studies, combined with computational analyses and novel approaches for structural determination, has allowed lately to slowly increase the amount of functional and structural information on these regions (Chavali et al. 2020; see a list of resources that gather information relevant to the study of polyQ regions in table 2).

We hope that in the future the further improvement of these experimental and computational resources will allow us to better explain the function of polyQ in human and other species. We could expect that such developments will similarly improve our knowledge on the function of other homorepeats and low complexity regions.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (AN735/4-1 to M.A.A.-N.).

Data Availability

The data underlying this article are available and in its online supplementary material.

Literature Cited

Albà MM, Santibáñez-Koref MF, Hancock JM. 2001. The comparative genomics of polyglutamine repeats: extreme differences in the codon organization of repeat-encoding regions between mammals and Drosophila. J Mol Evol. 52(3):249–259.

Albà MM, Guigó R. 2004. Comparative analysis of amino acid repeats in rodents and humans. Genome Res. 14(4):549–554.

Ashkenazi A, et al. 2017. Polyglutamine tracts regulate beclin 1-dependent autophagy. Nature 545(7652):108–111.

Atanesyan L, Günther V, Dichtl B, Georgiev O, Schaffner W. 2012. Polyglutamine tracts as modulators of transcriptional activation from yeast to mammals. Biol Chem. 393(1–2):63–70.

Bennett MJ, et al. 2002. A linear lattice model for polyglutamine in CAG-expansion diseases. Proc Natl Acad Sci U S A. 99(18):11634–11639.

Bhambri A, Pinto A, Pillai B. 2020. Interferon mediated neuroinflammation in polyglutamine disease is not caused by RNA toxicity. Cell Death Dis. 11(1):3.

Bhattacharyya A, et al. 2006. Oligoproline effects on polyglutamine conformation and aggregation. J Mol Biol. 355(3):524–535.

Buchanan LE, et al. 2014. Structural motif of polyglutamine amyloid fibrils discerned with mixed-isotope infrared spectroscopy. Proc Natl Acad Sci U S A. 111(16):5796–5801.

Chavali S, et al. 2017. Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. Nat Struct Mol Biol. 24(9):765–777.

Chavali S, Singh AK, Santhanam B, Babu MM. 2020. Amino acid homorepeats in proteins. Nat Rev Chem. 4(8):420–434.

Crick SL, Jayaraman M, Frieden C, Wetzel R, Pappu RV. 2006. Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions. Proc Natl Acad Sci U S A. 103(45):16764–16769.

Darnell G, Orgel JP, Pahl R, Meredith SC. 2007. Flanking polyproline sequences inhibit beta-sheet structure in polyglutamine segments by inducing PPII-like helix structure. J Mol Biol. 374(3):688–704.

Eftekharzadeh B, et al. 2016. Sequence context influences the structure and aggregation behavior of a polyQ tract. Biophys J. 110(11):2361–2366.

Escobedo A, et al. 2019. Side chain to main chain hydrogen bonds stabilize a polyglutamine helix in a transcription factor. Nat Commun. 10(1):2034.

Faber PW, et al. 1989. The N-terminal domain of the human androgen receptor is encoded by one, large exon. Mol Cell Endocrinol. 61(2):257–262.

Faux NG, et al. 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res. 15(4):537–551.

Fiumara F, Fioriti L, Kandel ER, Hendrickson WA. 2010. Essential role of coiled coils for aggregation and activity of Q/N-rich prions and PolyQ proteins. Cell 143(7):1121–1135.

Galzitskaya OV, Novikov GS, Dovidchenko NV, Lobanov MY. 2019. Is there codon usage bias for poly-Q stretches in the human proteome. J Bioinform Comput Biol. 17(1):1950010.

Gatchel JR, Zoghbi HY. 2005. Diseases of unstable repeat expansion: mechanisms and common principles. Nat Rev Genet. 6(10):743–755.

Gemayel R, et al. 2015. Variable glutamine-rich repeats modulate transcription factor activity. Mol Cell. 59(4):615–627.

Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium. 2019. CAG repeat not polyglutamine length determines timing of Huntington's disease onset. Cell 178(4):887–900.

Gerber HP, et al. 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 263(5148):808–811.

Hancock JM. 1996. Simple sequences and the expanding genome. Bioessays 18(5):421–425.

Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol. 4:2.

Holmes SE, et al. 1999. Expansion of a novel CAG trinucleotide repeat in the 5′ region of PPP2R2B is associated with SCA12. Nat Genet. 23(4):391–392.

Hosp F, et al. 2017. Spatiotemporal proteomic profiling of Huntington's disease inclusions reveals widespread loss of protein function. Cell Rep. 21(8):2291–2303.

Huntley M, Golding GB. 2000. Evolution of simple sequence in proteins. J Mol Evol. 51(2):131–140.

Jorda J, Kajava AV. 2010. Protein homorepeats sequences, structures, evolution, and functions. Adv Protein Chem Struct Biol. 79:59–88.

Kalathur RKR, Pedro Pinto J, Sahoo B, Chaurasia G, Futschik ME. 2017. HDNetDB: a molecular interaction database for network-oriented investigations into Huntington's disease. Sci Rep. 7(1):5216.

Karlin S, Burge C. 1996. Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. Proc Natl Acad Sci U S A. 93(4):1560–1565.

Klein FA, et al. 2007. Pathogenic and non-pathogenic polyglutamine tracts have similar structural properties: towards a length-dependent toxicity gradient. J Mol Biol. 371(1):235–244.

Komar AA. 2016. The Yin and Yang of codon usage. Hum Mol Genet. 25(R2):R77–R85.

Kuiper EFE, de Mattos EP, Jardim LB, Kampinga HH, Bergink S. 2017 Chaperones in polyglutamine aggregation: beyond the Q-stretch. Front Neurosci. 11(145):145.

Kwon MJ, et al. 2018. Coiled-coil structure-dependent interactions between polyQ proteins and Foxo lead to dendrite pathology and behavioral defects. Proc Natl Acad Sci U S A. 115(45):E10748–E10757.

La Spada AR, et al. 1992. Meiotic stability and genotype-phenotype correlation of the trinucleotide repeat in X-linked spinal and bulbar muscular atrophy. Nat Genet. 2(4):301–304.

Lavorgna G, Patthy L, Boncinelli E. 2001. Were protein internal repeats formed by "bricolage"? Trends Genet. 17(3):120–123.

Lee C, et al. 2013. Protein aggregation behavior regulates cyclin transcript localization and cell-cycle control. Dev Cell. 25(6):572–584.

Li LB, Yu Z, Teng X, Bonini NM. 2008. RNA toxicity is a component of ataxin-3 degeneration in Drosophila. Nature 453(7198):1107–1111.

Liu Y, Wilson SH. 2012. DNA base excision repair: a mechanism of trinucleotide repeat expansion. Trends Biochem Sci. 37(4):162–172.

Lobanov MY, Galzitskaya OV. 2012. Occurrence of disordered patterns and homorepeats in eukaryotic and bacterial proteomes. Mol Biosyst. 8(1):327–337.

Lobanov MY, Klus P, Sokolovsky IV, Tartaglia GG, Galzitskaya OV. 2016. Non-random distribution of homo-repeats: links with biological functions and human diseases. Sci Rep. 6:26941.

Lobanov MY, Sokolovskiy IV, Galzitskaya OV. 2014. HRaP: database of occurrence of HomoRepeats and patterns in proteomes. Nucleic Acids Res. 42(Database issue):D273–D278.

Lone WG, et al. 2016. Exploration of CAG triplet repeat in nontranslated region of SCA12 gene. J Genet. 95(2):427–432.

Masino L, Kelly G, Leonard K, Trottier Y, Pastore A. 2002. Solution structure of polyglutamine tracts in GST-polyglutamine fusion proteins. FEBS Lett. 513(2-3):267–272.

McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One. 8(4):e61217.

Mier P, Alanis-Lobato G, Andrade-Navarro MA. 2017. Context characterization of amino acid homorepeats using evolution, position, and order. Proteins 85(4):709–719.

Mier P, Andrade-Navarro MA. 2017. dAPE: a web server to detect homorepeats and follow their evolution. Bioinformatics 33(8):1221–1223.

Mier P, Andrade-Navarro MA. 2018. Glutamine codon usage and polyQ evolution in primates depend on the Q stretch length. Genome Biol Evol. 10(3):816–825.

Mier P, Andrade-Navarro MA. 2020. The features of polyglutamine regions depend on their evolutionary stability. BMC Evol Biol. 20(1):59.

Mier P, Elena-Real C, Urbanek A, Bernadó P, Andrade-Navarro MA. 2020. The importance of definitions in the study of polyQ regions: a tale of thresholds, impurities and sequence context. Comput Struct Biotechnol J. 18:306–313.

Nalavade R, Griesche N, Ryan DP, Hildebrand S, Krauss S. 2013. Mechanisms of RNA-induced toxicity in CAG repeat disorders. Cell Death Dis. 4:e752.

Pelassa I, et al. 2014. Association of polyalanine and polyglutamine coiled coils mediates expansion disease-related protein aggregation and dysfunction. Hum Mol Genet. 23(13):3402–3420.

Pelassa I, Fiumara F. 2015. Differential occurrence of interactions and interaction domains in proteins containing homopolymeric amino acid repeats. Front Genet. 6(345):345.

Perutz MF, Johnson T, Suzuki M, Finch JT. 1994. Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. Proc Natl Acad Sci U S A. 91(12):5355–5358.

Petrakis S, et al. 2012. Identification of human proteins that modify misfolding and proteotoxicity of pathogenic ataxin-1. PLoS Genet. 8(8):e1002897.

Petrakis S, Schaefer MH, Wanker EE, Andrade-Navarro MA. 2013. Aggregation of polyQ-extended proteins is promoted by interaction with their natural coiled-coil partners. Bioessays 35(6):503–507.

Ramazzotti M, Monsellier E, Kamoun C, Degl'Innocenti D, Melki R. 2012. Polyglutamine repeats are associated to specific sequence biases that are conserved among eukaryotes. PLoS One. 7(2):e30824.

Ross CA. 1997. Intranuclear neuronal inclusions: a common pathogenic mechanism for glutamine-repeat neurodegenerative diseases. Neuron 19(6):1147–1150.

Ruff KM, Khan SJ, Pappu RV. 2014. A coarse-grained model for polyglutamine aggregation modulated by amphipathic flanking sequences. Biophys J. 107(5):1226–1235.

Sayers EW, et al. 2009. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 37(Database issue):D5–D15.

Schaefer MH, Wanker EE, Andrade-Navarro MA. 2012. Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks. Nucleic Acids Res. 40(10):4273–4287.

Scherzinger E, et al. 1997. Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates in vitro and in vivo. Cell 90(3):549–558.

Shen K, et al. 2016. Control of the structural landscape and neuronal proteotoxicity of mutant Huntingtin by domains flanking the polyQ tract. Elife 5:e18065.

Siu HW, Heck B, Kovermann M, Hauser K. 2021. Template-assisted design of monomeric polyQ models to unravel the unique role of glutamine side chains in disease-related aggregation. Chem Sci. 12(1):412–426.

Srivastava AK, Takkar A, Garg A, Faruq M. 2017. Clinical behaviour of spinocerebellar ataxia type 12 and intermediate length abnormal CAG repeats in PPP2R2B. Brain 140(1):27–36.

Sumner-Smith M, Rafalski JA, Sugiyama T, Stoll M, Söll D. 1985. Conservation and variability of wheat alpha/beta-gliadin genes. Nucleic Acids Res. 13(11):3905–3916.

Szlachcic WJ, Switonski PM, Kurkowiak M, Wiatr K, Figiel M. 2015. Mouse polyQ database: a new online resource for research using mouse models of neurodegenerative diseases. Mol Brain. 8(1):69.

Totzeck F, Andrade-Navarro MA, Mier P. 2017. The protein structure context of polyQ regions. PLoS One. 12(1):e0170801.

The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47(D1):D506–D515.

Urbanek A, et al. 2020. Flanking regions determine the structure of the poly-glutamine in huntingtin through mechanisms common among glutamine-rich human proteins. Structure 28(7):733–746.e5.

Vázquez N, et al. 2019. EvoPPI 1.0: a web platform for within- and between-species multiple interactome comparisons and application to nine polyq proteins determining neurodegenerative diseases. Interdiscip Sci. 11(1):45–56.

Wharton KA, Yedvobnick B, Finnerty VG, Artavanis-Tsakonas S. 1985. opa: a novel family of transcribed repeats shared by the Notch locus and other developmentally regulated loci in *D. melanogaster*. Cell 40(1):55–62.

Whi Kim M, Chelliah Y, Woo Kim S, Otwinowski Z, Bezprozvanny I. 2009. Secondary structure of Huntingtin amino-terminal region. Structure 17(9):1205–1212.

Wright GEB, et al. 2019. Length of uninterrupted CAG, independent of polyglutamine size, results in increased somatic instability, hastening onset of Huntington disease. Am J Hum Genet. 104(6):1116–1126.

Zhao Y, et al. 2018. Comparative analysis of mutant huntingtin binding partners in yeast species. Sci Rep. 8(1):9554.

Zhemkov VA, Kulminskaya AA, Bezprozvanny IB, Kim M. 2016. The 2.2-Angstrom resolution crystal structure of the carboxy-terminal region of ataxin-3. FEBS Open Bio. 6(3):168–178.

Associate editor: Adam Eyre-Walker