Structure–function relationships in protein homorepeats

Carlos Elena-Real, Pablo Mier, Nathalie Sibille, Miguel Andrade-Navarro, Pau Bernadó

Structure-function relationships in protein homorepeats

Carlos A. Elena-Real,[a,#] Pablo Mier,[b,#] Nathalie Sibille,[a] Miguel A. Andrade-Navarro,[b] Pau Bernadó[a,*]

a- Centre de Biologie Structurale (CBS), Université de Montpellier, INSERM, CNRS. 29 rue de Navacelles, 34090 Montpellier, France.

b- Institute of Organismic and Molecular Evolution, Faculty of Biology, Johannes Gutenberg University Mainz. Hans-Dieter-Hüsch-Weg 15, 55128 Mainz, Germany.

# These authors have contributed equally to this work

*** Corresponding author:** Pau Bernadó (pau.bernado@cbs.cnrs.fr).

Tel.: +33 467417705

Other Authors' emails: C.A.E.R. (elena-real@cbs.cnrs.fr), P.M. (munoz@uni-mainz.de), N.S. (nathalie.sibille@cbs.cnrs.fr), M.A.A.N. (andrade@uni-mainz.de)

Short Title: Structure and Function in Homorepeats

Abstract

Homorepeats (or polyX), protein segments containing repetitions of the same amino acid, are abundant in proteomes from all kingdoms of life and are involved in crucial biological functions as well as several neurodegenerative and developmental diseases. Mainly inserted in disordered segments of proteins, the structure/function relationships of homorepeats remain largely unexplored. In this review, we summarize present knowledge for the most abundant homorepeats, highlighting the role of the inherent structure and the conformational influence exerted by their flanking regions. Recent experimental and computational methods enable residue-specific investigations of these regions and promise novel structural and dynamic information for this elusive group of proteins. This information should increase our knowledge about the structural bases of phenomena such as liquid-liquid phase separation and trinucleotide repeat disorders.

Highlights

- Proteins encompassing homorepeats (polyX) are abundant in all kingdoms of life
- Several diseases are caused by the expansion of existing homorepeats in proteins
- The nature of the amino acid defines the prevalent secondary structure of homorepeats
- Specific amino acid enrichments are observed for the majority of polyX flanking regions
- Site specific isotopic labelling and NMR provide unique information of homorepeats

Keywords: Homorepeat, intrinsically disordered protein, Site-specific Isotopic Labelling, Nuclear Magnetic Resonance

Introduction

A large percentage of protein sequences are aperiodic, showing close to average amino acid composition. This subtle mixture of residues dictates the structural properties of proteins and their functional role. However, there is an important group of proteins encompassing regions enriched in one or few amino acids, the so-called Low Complexity Regions (LCRs)[1*]. Mainly located within Intrinsically Disordered Regions (IDRs), *i.e.* regions without permanent secondary or tertiary structure[2,3], LCRs are found in half of eukaryotic proteins where they represent around 25% of the coding sequence[4]. Homorepeats (or polyX) are tracts of a single amino acid that represent an eye-catching family of LCRs[5–7]. Once considered as 'junk' protein segments without specific function, there is a growing body of evidence that underlines their biological relevance[1,8**]. Indeed, homorepeats exploit the accumulation of specific physicochemical properties in defined regions of proteins to perform very specialised functions in (among others) stress response, development, transcription, organelle biogenesis and transport[7,9]. Homorepeats provide functional versatility to proteins by mediating protein-protein interactions and driving spatial localization[8**,10]. Moreover, their presence, even in essential proteins, facilitates protein divergence and evolvability to rewire interactions[11,12]. It has been also shown that proteins containing homorepeats have denser and more diverse interactomes[8], and these containing multiple polyX are more often involved in disease, including neurological disorders and cancer[13,14]. Although protein length is a factor that needs to be taken into account because it necessarily increases the probability to find more polyX, these two observations underline the role of homorepeats in signalling and regulatory processes.

The accumulation of a given physicochemical feature can also have detrimental consequences. Indeed, repeats of certain amino acids, such as cysteine, tyrosine or tryptophan, are rarely found in proteomes, suggesting their inherent toxicity. Moreover, the uncontrolled expansion of poly-glutamine (polyQ) and poly-alanine (polyA) in specific proteins cause a series of rare neurodegenerative and developmental diseases, including Huntington's disease, several ataxias, synpolydactyly syndrome and Ondine's curse [15–18]. These pathologies are triggered by the incorporation of few additional residues in a previously existing homorepeat, demonstrating the subtle balance between

function and toxicity[19]. More recently, polyG aggregates originating from expanded $(CGG)_n$ repeats located in 5'-untranslated regions of certain genes have been identified in patients of (among others) neural intranuclear inclusion disease and fragile X tremor/ataxia syndrome[20**,21].

Despite the growing attention to homorepeats, the structural bases of their function and malfunction remain poorly understood, precluding rational intervention for biomedical purposes. Moreover, a precise control of the structural determinants of these sequences would pave the way to design of IDRs with targeted functions in biotechnology[22]. LCRs in general and homorepeats in particular pose fundamental problems for the application of traditional high-resolution structural biology methods. On the one hand, their inherent flexibility precludes the general use of X-ray crystallography and cryo-electron microscopy. On the other hand, the similarity of the chemical environments in repetitive sequences hampers the application of standard Nuclear Magnetic Resonance (NMR) frequency assignment strategies[23]. These limitations have fostered the application of low-resolution methods and computational approaches in order to establish connections between the structure of homorepeats and their biological function[24–26]. Complementary to these methods, computational and genomic approaches have been applied to assess the distribution of homorepeats in proteomes and to evaluate their interactome and evolutionary dynamics [27–30**].

In this review, we summarize present structural knowledge for the most abundant homorepeats and describe recent developments to study the structure/function relationships of this elusive group of proteins.

Abundance of homorepeats

Several studies have surveyed the abundance of polyX in various organisms, finding high variability between species in frequency and type of polyX[10,31,32]. When assessing this variability, the definition of the motif used for identification (minimal length and allowed non-X residues) is critical, as it changes the sensitivity in detecting functionally relevant polyX[33]. While several studies indicate that eukaryota tend to have more polyX than prokaryota, there is no trend associating a larger content of polyX with particular organismic properties (like being multicellular). For instance, there are DNA viruses, such as Pandoraviruses[34], that have more polyX than many non-viral species

(10% of their proteins contain at least one homorepeat), and some unicellular organisms have the highest content of polyN. For example, *Dictyostelium discoideum* and *Plasmodium falciparum* have 51% and 56% of their proteins with at least one polyN, respectively[35]. The ensemble of these surveys indicates that, despite the fact that polyX are widely distributed among species, their enrichment in functional classes is not necessarily conserved[6].

Intrinsic conformational preferences of polyX

The chemical nature of the repeated amino acid is the main determinant of the preferred secondary structure of a polyX. Systematic surveys of polyX fragments in the PDB have identified some of these conformational preferences[1,6,36–38]. Although insightful in some cases, the low abundance of polyX in crystallographic structures, their limited length and the strong influence exerted by the flanking regions (see below) limit the generalization of these observations. Using the LS2P server[39], which quantifies the structural variability of concatenated tripeptides derived from high-resolution structures, the intrinsic conformational preferences for the 20 polyX were predicted (Figure 1). PolyX can be structurally classified in four groups. *(i)* A, C, E and L have a strong tendency to adopt α-helical conformations. *(ii)* Extended (β-strand and PP-II) conformations are preferred for H, I, P, V, W and Y, mainly due to the bulkiness of the side chains that hamper compact helical conformations[40]. *(iii)* Homorepeats of D, G, M, N and S mainly adopt non-canonical conformations. *(iv)* Finally, F, K, Q, R and T homorepeats exhibit non-negligible populations of canonical and non-canonical conformations. The plasticity observed in these two last groups is interpreted as a sign of structural disorder and the possibility to be conformationally influenced by flanking sequences and the environment. Interestingly, chemically similar residues can display different behaviour. For instance, homorepeats of the two acidic amino acids, D and E, belong to two different groups. Similarly, polyL seems to prefer α-helical secondary structure, in contrast to also hydrophobic polyI and polyV, which prefer the extended one. Finally, while positive amino acids K and R have a ~15-20% tendency for extended conformations, polyH shows a 92%.

Influence of sequence context in the structure of homorepeats

The second main factor governing the structure of a homorepeat is its sequence context. The conformation of the residues flanking a homorepeat can propagate in it, or can alter the intrinsic structural properties of the homorepeat, especially if it is short[27,41]. Only polyQ and polyA have been analyzed in detail in this regard[33,37,42,43]. In both cases, specific amino acid enrichments in flanking regions were observed. In polyQ this enrichment was asymmetric, with L and P overrepresented in the N- and C-flanks of the homorepeat, respectively[33,42,43]. Conversely, a symmetric enrichment in P and G was found in polyA flanking regions[37].

Here, we have extended the bioinformatic sequence context study to all different homorepeats in the human proteome. Homorepeats were identified using the polyX2 tool in standalone mode with a lax threshold of a minimum of 4 identical residues in a window of 6 amino acids[44*]. Then, we calculated the amino acid abundance per position in the five N- and C-terminal amino acids surrounding each homorepeat (Figure 2). Results show that most polyX display an enrichment of the amino acid X in the vicinity, with the exception of polyL, polyV and polyS. In other words, most polyX types are located in X-rich regions. Interestingly, enrichments of amino acids different from the one of the homorepeat are also observed. For instance, positions around polyD are highly enriched in E and, conversely, positions around polyE are highly enriched in D, pointing to highly charged DE-rich protein regions[45]. We observe a similar association for S and T. The capacity of both amino acids to be phosphorylated could strongly modify the structural and functional properties of these regions upon external stimuli. Flanking regions of polyK are enriched in E and R, giving rise to highly charged protein stretches. Some amino acids appear enriched in the flanking regions of multiple polyX types. For instance, G is found specially enriched in polyA, polyP and polyS flanks, while P is found close to polyA, polyG, polyQ, polyS and polyT.

PolyQ as the prototypical example of homorepeat

PolyQ is one of the most abundant homorepeats in eukaryotes. Computational analysis of their sequence context in proteins and some experimental evidence suggest that polyQ have a function in extending the conformation of an adjacent N-terminal coiled coil region upon its interaction with the coiled coil of another protein[19,46]. This would

explain why genetic mutations changing the length of the polyQ could affect their interactome resulting in pathogenic interactions and aggregates [30,47]. Indeed, nine inherited human diseases have been lined to aberrant expansion of polyQ and subsequent amyloid formation[17]. In these pathologies, the polyQ length and the number of consecutive CAG codons are correlated with the propensity to aggregate *in vitro* and to form inclusions in neurons, as well as the disease severity and age of onset[17,48]. As a consequence, polyQ is the most studied homorepeat from a functional and structural perspective[49]. Despite these efforts, contradictory structural models, mainly based on sparse or low-resolution data, have been proposed to explain the molecular mechanism of pathogenicity[26,50–53].

Recent works have shed light into the main parameters governing the structure of polyQ and the influence of their flanking regions. NMR investigations of an N-terminal fragment of the androgen receptor (AR) and the exon-1 of huntingtin (HttExon-1) have demonstrated the presence of hydrogen bonds between both the backbone and the side chain amines of glutamines in position *i* with residues located in position *i-4* (Figure 3A)[42,54**]. These bifurcated hydrogen bonds propagate and stabilise the α-helical structure along the polyQ. Interestingly, the four leucines and the phenylalanine preceding the AR and the HttExon-1 polyQ, respectively, are key to the stabilisation of these hydrogen bonds and consequently of the helical structure. Indeed, by mutating residues in the flanking regions one can control and modulate the helical propensity of polyQ tracts (Figure 3B)[42,54**]. In order to rationalize these observations, a systematic investigation of the effect of N-flanking residues on the structure of AR polyQ was performed, showing that large and hydrophobic residues in this position (W, Y, F, I, V and L) strongly stabilize the α-helical propensity[55]. The role of α-helix stability in the aggregation propensity is, however, unclear. While in AR the helical destabilisation seems to enhance aggregation[56], the inverse effect has been reported for HttExon-1[57**]. Importantly, the enrichment of L in the proximity of polyQ tracts and glutamine-rich proteins in eukaryotes suggests that this is a general evolutionary conserved structural mechanism[33].

The recent development of Site-Specific Isotopic Labelling (SSIL)[58], which combines the tRNA suppressor strategy and cell free protein expression, is a promising methodology to derive new high-resolution structural information of homorepeats in a

length-independent manner[23]. SSIL has enabled to resolve highly degenerated NMR spectra of non-pathogenic (16Q)[42], within the threshold (36Q)[59] and pathogenic (46 and 66Q)[57**] versions of HttExon-1 to obtain structural information that was out of reach for traditional approaches. The combined analyses of NMR data with Small Angle X-ray Scattering and molecular dynamics simulations have shown the concomitant increase of helicity upon polyQ expansion through the presence of bifurcated hydrogen bonds (Figure 3C). This gradual effect is less apparent in AR, probably due to the lack of structural information for pathogenic versions of this protein. Importantly, the increase in helicity has been shown to be a key element in the aggregation propensity of HttExon-1 and other polyQ-hosting proteins, probably by enhancing the formation of productive dimers, tetramers and other oligomers through coiled-coil interactions[60–62] that eventually can phase separate[63,64] and/or produce inclusion bodies in neurons [65,66*].

Connecting structure and function for other homorepeats

The structural knowledge accumulated for homorepeats different than polyQ is more limited. In this section, we have compiled the present structural understanding for some of them.

PolyA has attracted a great deal of attention because there are eight developmental diseases caused by the abnormal expansion of this homorepeat in several transcription factors[15,16]. Successive studies of short capped peptides provided contradictory evidence indicating either a strong propensity to adopt α-helical structure[67] or to be disordered with some prevalence for extended poly-proline-II conformations[68,69]. However, when polyA is studied in its protein context, a α-helical propensity has been reported[22,70,71]. A recent SSIL study of the two polyA tracts of Phox2B, containing 9 and 20 alanines, confirmed the helical tendency for this homorepeat, and highlighted the correlation between polyA length and conformational stability, suggesting the presence of cooperative effects[72]. This last feature can explain some of the functional observations of alanine-rich proteins, including the enhanced aggregation and phase separation propensities of expanded polyA in disease-related proteins[73,74]. Similarly to polyQ, upon expansion, longer and more stable polyA α-helices could favor coiled-coil intramolecular interactions[75**,76]. Indeed, the enrichment in alanine residues has

been observed in several frameshift mutations inducing genetic diseases associated with misregulation of phase-separation phenomena[77].

PolyP regions participate in protein-protein interaction networks, often through specific interactions with domains such as SRC homology 3 (SH3) and the WW[78,79]. However, profilins, small actin-binding proteins, are the only reported example requiring at least 6-8 consecutive prolines for high affinity binding[80,81]. In aqueous solutions, polyP adopt extended rod-like helix known as poly-proline type-II structure that is stabilised by $n\text{->}\pi^*$ interactions between adjacent carbonyl groups ($C_{i-1}=O_{i-1}...C_i=O_i$)[82–84]. Due to its cyclic nature, proline is the only amino acid that presents *cis/trans* isomeric equilibrium in noticeable amounts to be detected experimentally. Whether this capacity is maintained within polyP tracts remained an open question. Using smFRET, it has been shown that, in long polyP, the *cis* population of inner prolines is severely reduced (≈2%) with respect to these positioned at the termini (≈10%)[24]. This has been recently confirmed in a NMR study using SSIL samples, where the cooperativity of $n\text{->}\pi^*$ interactions was suggested as the origin of the reduced *cis* population for inner prolines[85*]. The inherent stiffness of polyP has been associated with its protective role in the C-terminus of aggregation-prone polyQ, being HttExon-1 the most notorious example[41,42,86]. Indeed, the coevolution of both homorepeats has been suggested as general mechanism where polyP emerge in evolution after a polyQ has been established [19,42,87,88]. Moreover, the structural rigidity of polyP has been exploited to design molecular rulers and scaffolds for bioengineering applications[89,90].

In line with our bioinformatics analysis (Figure 2), polyS-containing peptides have been described to adopt distinct conformations depending on the sequence context[91]. In addition to their role in protein localization and regulation of phase separation[92,93*,94], polyS can also emerge through frameshift or repeat-associated non-AUG translation of polyQ tracts, eventually contributing to the overall toxicity of the expanded gene[95,96]. In a recent study, it has been shown that in many cases polyS originating from aberrant translation adopt helical structures with a high propensity to form coiled-coils, which promote the oligomerization and fibrilization *in vitro*[91]. In this sense, polyS behaves similarly to previously described polyQ and polyA.

Concluding remarks

There is a growing body of evidence showing that protein homorepeats perform multiple pivotal biological functions in all kingdoms of life. As these observations originate mainly from functional and computational studies, the link between the biophysical and structural features of homorepeats with their functions is, in the vast majority of cases, poorly understood. In recent years novel NMR strategies based on ^{13}C-detected experiments have demonstrated their power to resolve highly crowded spectra and provide residue-specific information for relatively long polyP and polyQ, especially when these tracts exhibit high levels of structuration[56,97,98]. The capacity to isotopically enrich individual residues provided by the SSIL strategy renders high-resolution NMR studies independent of the length and level of structuration of polyX stretches. At present, only glutamine[58], proline[85*] and alanine[72] are available for specific labelling, but the extension to other amino acids should make SSIL a universal tool to study polyX stretches. In the absence of extensive structural data for homorepeats, computational structural biology approaches represent an excellent alternative. Great efforts have been done in the recent years to deliver robust protein force-fields and water models to derive atomistic models of disordered proteins with the capacity to reproduce experimental data[99–102]. The refinement of these physical models to accurately simulate LCRs and homorepeats is a logical subsequent step[103*]. However, this improvement can only be robustly achieved if enough experimental data on these systems is made available to computational scientists. In this context, the two most complete structural studies of the polyQ proteins have combined residue-specific NMR data with MD simulations to unveil the atomistic details governing the conformation of this homorepeat, highlighting the power of this combination[54**,57**]. The characterization of intermolecular interactions involving homorepeats giving rise to aggregation or phase separation represents a crucial challenge for the future[104]. From a computational perspective, these studies are normally addressed by using simplified coarse-grained protein models, whose parametrization will also depend on the availability of accurate experimental data[105–107].

In summary, uncovering the role of structure in defining complex functions of the different homorepeats will require the combination of low-resolution and residue-specific information from experimental methods in combination with accurate

computational approaches and large-scale bioinformatics studies. We believe that only the application of these integrative strategies will bring light to the fraction of the "dark proteome"[108] represented by homorepeats.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 778247 and No 823886. This publication is based upon work from COST Action ML4NGP, CA21160, supported by COST (European Cooperation in Science and Technology). This work was also supported by the European Research Council under the European Union's H2020 Framework Programme (2014-2020) / ERC Grant agreement n° [648030], Labex EpiGenMed, an "Investissements d'avenir" program (ANR-10-LABX-12-01), and MUSE-App 2021 (Ondine) awarded to PB. The CBS is a member of France-BioImaging (FBI) and the French Infrastructure for Integrated Structural Biology (FRISBI), two national infrastructures supported by the French National Research Agency (ANR-10-INBS-04-01 and ANR-10-INBS-05, respectively).

Conflict of Interest

The authors declare no competing interest.

References

* of special interest

** of outstanding interest

1*. Mier P, Paladin L, Tamana S, Petrosian S, Hajdu-Soltész B, Urbanek A, Gruca A, Plewczynski D, Grynberg M, Bernadó P, et al.: **Disentangling the complexity of low complexity proteins**. *Brief Bioinform* 2020, **21**:458–472.

In this review and using muliple protein sequences as example, the authors clarify the definitions and relationships of concepts such as low-complexity region, compositional bias and homorepeat.

2. Van Der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, et al.: **Classification of intrinsically disordered regions and proteins**. *Chem Rev* 2014, **114**:6589–6631.

3. Wright PE, Dyson HJ: **Intrinsically disordered proteins in cellular signalling**

and regulation. *Nat Rev Mol Cell Biol* 2015, **16**:18–29.

4. Wootton JC: **Non-globular domains in protein sequences: Automated segmentation using complexity measures**. *Comput Chem* 1994, **18**:269–285.

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

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

7. Jorda J, Kajava A V.: **Protein homorepeats: Sequences, structures, evolution, and functions**. In *Advances in Protein Chemistry and Structural Biology*. . 2010:59–88.

8**. Chavali S, Chavali PL, Chalancon G, De Groot NS, Gemayel R, Latysheva NS, Ing-Simmons E, Verstrepen KJ, Balaji S, Babu MM: **Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins**. *Nat Struct Mol Biol* 2017, **24**:765–777.

In this outstanding study, the authors address the reasons for the enrichment of homorepeats in eukaryota despite their potential detrimental effects. They show that homorepeats are enriched in homeostatic proteins and that they increase the functional versatility of proteins facilitating new interactions and spatial organization. Through mutations, homorepeats enable the rapid exploration of the genotype-phenotype landscape, contributing to adaptation.

9. Lobanov MY, Klus P, Sokolovsky I V., Tartaglia GG, Galzitskaya O V.: **Non-random distribution of homo-repeats: Links with biological functions and human diseases**. *Sci Rep* 2016, **6**:26941.

10. Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, De La Banda MG, Whisstock JC: **Functional insights from the distribution and role of homopeptide repeat-containing proteins**. *Genome Res* 2005, **15**:537–551.

11. Gemayel R, Chavali S, Pougach K, Legendre M, Zhu B, Boeynaems S, van der Zande E, Gevaert K, Rousseau F, Schymkowitz J, et al.: **Variable Glutamine-Rich Repeats Modulate Transcription Factor Activity**. *Mol Cell* 2015, **59**:615–627.

12. Singh AK, Amar I, Ramadasan H, Kappagantula KS, Chavali S: **Proteins with amino acid repeats constitute a rapidly evolvable and human-specific essentialome**. *Cell Rep* 2023, **42**:112811.

13. Karlin S, Brocchieri L, Bergman A, Mrázek J, Gentles AJ: **Amino acid runs in eukaryotic proteomes and disease associations**. *Proc Natl Acad Sci U S A* 2002, **99**:333–338.

14. Meszaros A, Ahmed J, Russo G, Tompa P, Lazar T: **The evolution and polymorphism of mono-amino acid repeats in androgen receptor and their regulatory role in health and disease**. *Front Med* 2022, **9**.

15. Amiel J, Trochet D, Clément-Ziza M, Munnich A, Lyonnet S: **Polyalanine expansions in human**. *Hum Mol Genet* 2004, **13**:R235–R243.

16. Darling AL, Uversky VN: **Intrinsic Disorder in Proteins with Pathogenic Repeat Expansions.** *Molecules* 2017, **22**:2027.

17. Stoyas CA, La Spada AR: **The CAG–polyglutamine repeat diseases: a clinical, molecular, genetic, and pathophysiologic nosology**. In *Handbook of Clinical Neurology*. . 2018:143–170.

18. Shao J, Diamond MI: **Polyglutamine diseases: Emerging concepts in pathogenesis and therapy**. *Hum Mol Genet* 2007, **16**:R115–R123.

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

20**. Sellier C, Buijsen RAM, He F, Natla S, Jung L, Tropel P, Gaucherot A, Jacobs H, Meziane H, Vincent A, et al.: **Translation of Expanded CGG Repeats into FMRpolyG Is Pathogenic and May Contribute to Fragile X Tremor Ataxia Syndrome**. *Neuron* 2017, **93**:331–347.

In this study, the authors show that the expanded $(CGG)_n$ repeats associated to neurodegenerative fragile X-associated tremor/ataxia syndrome (FXTAS) can be translated to polyG-containing proteins. They also show that these expanded homorepeats are pathogenic and form inclusions in neurons. This study has contributed to the description of a the new family of diseases linked to the polyG expansion.

21. Liufu T, Zheng Y, Yu J, Yuan Y, Wang Z, Deng J, Hong D: **The polyG diseases: a new disease entity**. *Acta Neuropathol Commun* 2022, **10**:79.

22. Roberts S, Harmon TS, Schaal JL, Miao V, Li K (Jonathan), Hunt A, Wen Y, Oas TG, Collier JH, Pappu R V., et al.: **Injectable tissue integrating networks from recombinant polypeptides with tunable order**. *Nat Mater* 2018, **17**:1154–1163.

23. Urbanek A, Elena-Real CA, Popovic M, Morató A, Fournet A, Allemand F, Delbecq S, Sibille N, Bernadó P: **Site-Specific Isotopic Labeling (SSIL): Access to High-Resolution Structural and Dynamic Information in Low-Complexity Proteins**. *ChemBioChem* 2020, **21**:769–775.

24. Best RB, Merchant KA, Gopich I V., Schuler B, Bax A, Eaton WA: **Effect of flexibility and cis residues in single-molecule FRET studies of polyproline**. *Proc Natl Acad Sci U S A* 2007, **104**:18964–18969.

25. Greving I, Dicko C, Terry A, Callow P, Vollrath F: **Small angle neutron scattering of native and reconstituted silk fibroin**. *Soft Matter* 2010, **6**:4389–4395.

26. Bravo-Arredondo JM, Kegulian NC, Schmidt T, Pandey NK, Situ AJ, Ulmer TS, Langen R: **The folding equilibrium of huntingtin exon 1 monomer depends on its polyglutamine tract.** *J Biol Chem* 2018, **293**:19613–19623.

27. Chiu SH, Ho WL, Sun YC, Kuo JC, Huang J rong: **Phase separation driven by interchangeable properties in the intrinsically disordered regions of protein paralogs**. *Commun Biol* 2022, **5**:400.

28. Lavoie H, Debeane F, Trinh QD, Turcotte JF, Corbeil-Girard LP, Dicaire MJ, Saint-Denis A, Pagé M, Rouleau GA, Brais B: **Polymorphism, shared functions and convergent evolution of genes with sequences coding for polyalanine domains**. *Hum Mol Genet* 2003, **12**:2967–2979.

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

30**. Vaglietti S, Fiumara F: **PolyQ length co-evolution in neural proteins**. *NAR Genomics Bioinforma* 2021, **3**:lqab032.

In this study, the authors explore the evolution of polyQ length and its co-evolution across functionally related protein pairs and networks. They detect an evolutionary hypervariability of polyQ length in neural proteins as well as an extensive co-variation of the polyQ length in interacting proteins.

31. Lobanov MY, Galzitskaya O V: **Occurrence of disordered patterns and homorepeats in eukaryotic and bacterial proteomes.** *Mol Biosyst* 2012, **8**:327–337.

32. Lobanov MY, Sokolovskiy I V., Galzitskaya O V.: **HRaP: Database of occurrence of HomoRepeats and patterns in proteomes**. *Nucleic Acids Res* 2014, **42**.

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

34. Erdozain S, Barrionuevo E, Ripoll L, Mier P, Andrade-Navarro MA: **Protein repeats evolve and emerge in giant viruses**. *J Struct Biol* 2023, **215**:107962.

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

36. Totzeck F, Andrade-Navarro MA, Mier P: **The protein structure context of polyQ regions**. *PLoS One* 2017, **12**.

37. Mier P, Elena-Real CA, Cortés J, Bernadó P, Andrade-Navarro MA: **The sequence context in poly-alanine regions: structure, function and conservation**. *Bioinformatics* 2022, **38**:4851–4858.

38. Gonçalves-Kulik M, Mier P, Kastano K, Cortés J, Bernadó P, Schmid F, Andrade-Navarro MA: **Low Complexity Induces Structure in Protein Regions Predicted as Intrinsically Disordered**. *Biomolecules* 2022, **12**.

39. Estaña A, Barozet A, Mouhand A, Vaisset M, Zanon C, Fauret P, Sibille N, Bernadó P, Cortés J: **Predicting Secondary Structure Propensities in IDPs Using Simple Statistics from Three-Residue Fragments**. *J Mol Biol* 2020, **432**:5447–5459.

40. Cho M-K, Kim H-Y, Bernado P, Fernandez CO, Blackledge M, Zweckstetter M: **Amino acid bulkiness defines the local conformations and dynamics of natively unfolded alpha-synuclein and tau.** *J Am Chem Soc* 2007, **129**:3032–3.

41. Bhattacharyya A, Thakur AK, Chellgren VM, Thiagarajan G, Williams AD, Chellgren BW, Creamer TP, Wetzel R: **Oligoproline effects on polyglutamine conformation and aggregation**. *J Mol Biol* 2006, **355**:524–535.

42. Urbanek A, Popovic M, Morató A, Estaña A, Elena-Real CA, Mier P, Fournet A, Allemand F, Delbecq S, Andrade-Navarro MA, et al.: **Flanking Regions Determine the Structure of the Poly-Glutamine in Huntingtin through Mechanisms**

Common among Glutamine-Rich Human Proteins. *Structure* 2020, **28**:733–746.e5.

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

44*. Mier P, Andrade-Navarro MA: **PolyX2: Fast Detection of Homorepeats in Large Protein Datasets**. *Genes (Basel)* 2022, **13**:758.

In this study, the authors present PolyX2, a powerful tool for the fast and efficient search for homorepeats in protein datasets. The nature of the homorepeat, the length of the scanned window and its purity can be tuned, making PolyX2 a very versatile tool for homorepeat investigations.

45. Bigman LS, Iwahara J, Levy Y: **Negatively Charged Disordered Regions are Prevalent and Functionally Important Across Proteomes**. *J Mol Biol* 2022, **434**:167660.

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

47. Mier P, Andrade-Navarro MA: **Between Interactions and Aggregates: The PolyQ Balance**. *Genome Biol Evol* 2021, **13**:evab246.

48. Lee JM, Correia K, Loupe J, Kim KH, Barker D, Hong EP, Chao MJ, Long JD, Lucente D, Vonsattel JPG, et al.: **CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset**. *Cell* 2019, **178**:887–900.e14.

49. Barbosa Pereira PJ, Manso JA, Macedo-Ribeiro S: **The structural plasticity of polyglutamine repeats**. *Curr Opin Struct Biol* 2023, **80**:102607.

50. Warner JB, Ruff KM, Tan PS, Lemke EA, Pappu R V, Lashuel HA: **Monomeric Huntingtin Exon 1 Has Similar Overall Structural Features for Wild-Type and Pathological Polyglutamine Lengths.** *J Am Chem Soc* 2017, **139**:14456–14469.

51. Kang H, Vázquez FX, Zhang L, Das P, Toledo-Sherman L, Luan B, Levitt M, Zhou R: **Emerging β-Sheet Rich Conformations in Supercompact Huntingtin Exon-1 Mutant Structures**. *J Am Chem Soc* 2017, **139**:8820–8827.

52. Moldovean SN, Chiş V: **Molecular Dynamics Simulations Applied to Structural and Dynamical Transitions of the Huntingtin Protein: A Review**. *ACS Chem Neurosci* 2020, **11**:105–120.

53. Feng X, Luo S, Lu B: **Conformation Polymorphism of Polyglutamine Proteins.** *Trends Biochem Sci* 2018, **43**:424–435.

54**. Escobedo A, Topal B, Kunze MBA, Aranda J, Chiesa G, Mungianu D, Bernardo-Seisdedos G, Eftekharzadeh B, Gairí M, Pierattelli R, et al.: **Side chain to main chain hydrogen bonds stabilize a polyglutamine helix in a transcription factor.** *Nat Commun* 2019, **10**:2034.

By combining NMR and computational methodologies, the authors have identified bifurcated hydrogen bonds involving side chain and backbone amide groups of glutamines as the structural element inducing helicity in androgen receptor polyQ

tract. They demonstrate that leucines placed at the N-flaking region of the polyQ stabilize bifurcate hydrogen bonds.

55. Escobedo A, Piccirillo J, Aranda J, Diercks T, Mateos B, Garcia-Cabau C, Sánchez-Navarro M, Topal B, Biesaga M, Staby L, et al.: **A glutamine-based single α-helix scaffold to target globular proteins**. *Nat Commun* 2022, **13**:7073.

56. Eftekharzadeh B, Piai A, Chiesa G, Mungianu D, García J, Pierattelli R, Felli IC, Salvatella X: **Sequence Context Influences the Structure and Aggregation Behavior of a PolyQ Tract.** *Biophys J* 2016, **110**:2361–2366.

57**. Elena-Real CA, Sagar A, Urbanek A, Popovic M, Morató A, Estaña A, Fournet A, Doucet C, Lund XL, Shi ZD, et al.: **The structure of pathogenic huntingtin exon 1 defines the bases of its aggregation propensity**. *Nat Struct Mol Biol* 2023, **30**:309–320.

Residue-specific NMR investigation of pathogenic HttExon-1 constructs containing 46 and 66 consecutive glutamines using SSIL samples. The authors show the persistence of the α-helical structure in pathogenic huntingtin, which is propagated and stabilized by bifurcated hydrogen bonds. Using mutants of the N-flanking region, they correlate the helical stability with the capacity to aggregate *in vitro* and to form inclusions in cell.

58. Urbanek A, Morató A, Allemand F, Delaforge E, Fournet A, Popovic M, Delbecq S, Sibille N, Bernadó P: **A General Strategy to Access Structural Information at Atomic Resolution in Polyglutamine Homorepeats.** *Angew Chem Int Ed Engl* 2018, **57**:3598–3601.

59. Elena-Real CA, Urbanek A, Lund XL, Morató A, Sagar A, Fournet A, Estaña A, Bellande T, Allemand F, Cortés J, et al.: **Multi-site-specific isotopic labeling accelerates high-resolution structural investigations of pathogenic huntingtin exon-1**. *Structure* 2023, **31**:644–650.e5.

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

61. Kotler SA, Tugarinov V, Schmidt T, Ceccon A, Libich DS, Ghirlando R, Schwieters CD, Clore GM: **Probing initial transient oligomerization events facilitating Huntingtin fibril nucleation at atomic resolution by relaxation-based NMR.** *Proc Natl Acad Sci U S A* 2019, **116**:3562–3571.

62. Ceccon A, Tugarinov V, Clore GM: **Quantitative Exchange NMR-Based Analysis of Huntingtin-SH3 Interactions Suggests an Allosteric Mechanism of Inhibition of Huntingtin Aggregation**. *J Am Chem Soc* 2021, **143**:9672–9681.

63. Peskett TR, Rau F, O'Driscoll J, Patani R, Lowe AR, Saibil HR: **A Liquid to Solid Phase Transition Underlying Pathological Huntingtin Exon1 Aggregation**. *Mol Cell* 2018, **70**:588–601.e6.

64. Hutin S, Kumita JR, Strotmann VI, Dolata A, Ling WL, Louafi N, Popov A, Milhiet P-E, Blackledge M, Nanao MH, et al.: **Phase separation and molecular ordering of the prion-like domain of the Arabidopsis thermosensory protein EARLY**

FLOWERING 3. *Proc Natl Acad Sci* 2023, **120**:e2304714120.

65. Bäuerlein FJB, Saha I, Mishra A, Kalemanov M, Martínez-Sánchez A, Klein R, Dudanova I, Hipp MS, Hartl FU, Baumeister W, et al.: **In Situ Architecture and Cellular Interactions of PolyQ Inclusions**. *Cell* 2017, **171**:179–187.e10.

66*. Riguet N, Mahul-Mellier AL, Maharjan N, Burtscher J, Croisier M, Knott G, Hastings J, Patin A, Reiterer V, Farhan H, et al.: **Nuclear and cytoplasmic huntingtin inclusions exhibit distinct biochemical composition, interactome and ultrastructural properties**. *Nat Commun* 2021, **12**:6579.

Using electron tomography the authors investigate the structure of cytosolic and nuclear inclusions of HttExon-1 in mammalian cells and primary neurons. They investigate the complex process of maturation and identify organelles, membranes and proteins sequestered by these inclusions.

67. Gratzer WB, Doty P: **A Conformation Examination of Poly-L-alanine and Poly-D,L-alanine in Aqueous Solution**. *J Am Chem Soc* 1963, **85**:1193–1197.

68. Chen K, Liu Z, Kallenbach NR: **The polyproline II conformation in short alanine peptides is noncooperative**. *Proc Natl Acad Sci U S A* 2004, **101**:15352–15357.

69. Shi Z, Anders Olson C, Rose GD, Baldwin RL, Kallenbach NR: **Polyproline II structure in a sequence of seven alanine residues**. *Proc Natl Acad Sci U S A* 2002, **99**:9190–9195.

70. Chen T-C, Huang J: **Musashi-1: An Example of How Polyalanine Tracts Contribute to Self-Association in the Intrinsically Disordered Regions of RNA-Binding Proteins**. *Int J Mol Sci* 2020, **21**.

71. Hong JY, Wang DD, Xue W, Yue HW, Yang H, Jiang LL, Wang WN, Hu HY: **Structural and dynamic studies reveal that the Ala-rich region of ataxin-7 initiates α-helix formation of the polyQ tract but suppresses its aggregation**. *Sci Rep* 2019, **9**:7481.

72. Elena-Real CA, Urbanek A, Imbert L, Morató A, Fournet A, Allemand F, Sibille N, Boisbouvier J, Bernadó P: **Site-Specific Introduction of Alanines for the NMR Investigation of Low-Complexity Regions and Large Biomolecular Assemblies**. *bioRxiv* 2023, doi:10.1101/2023.05.08.539737.

73. Polling S, Ormsby AR, Wood RJ, Lee K, Shoubridge C, Hughes JN, Thomas PQ, Griffin MDW, Hill AF, Bowden Q, et al.: **Polyalanine expansions drive a shift into α-helical clusters without amyloid-fibril formation**. *Nat Struct Mol Biol* 2015, **22**:1008–1015.

74**. Basu S, Mackowiak SD, Niskanen H, Knezevic D, Asimi V, Grosswendt S, Geertsema H, Ali S, Jerković I, Ewers H, et al.: **Unblending of Transcriptional Condensates in Human Repeat Expansion Disease**. *Cell* 2020, **181**:1062–1079.e30.

Using HOXD13 transcription factor as example, the authors demonstrate that the expansion of the polyA tract, which causes hereditary synpolydactyly in humans, perturbs the capacity of the protein to phase separate *in vitro* and *in vivo*, modifies the composition of the condensates, and alters the transcriptional program. These observations were generalized to other disease-associated polyA expansions in other transcription factors.

75. Nojima J, Oma Y, Futai E, Sasagawa N, Kuroda R, Turk B, Ishiura S: **Biochemical analysis of oligomerization of expanded polyalanine repeat proteins**. *J Neurosci Res* 2009, **87**:2290–2296.

76. Pelassa I, Corà D, Cesano F, Monje FJ, Montarolo PG, Fiumara F: **Association of polyalanine and polyglutamine coiled coils mediates expansion disease-related protein aggregation and dysfunction**. *Hum Mol Genet* 2014, **23**:3402–3420.

77. Mensah MA, Niskanen H, Magalhaes AP, Basu S, Kircher M, Sczakiel HL, Reiter AMV, Elsner J, Meinecke P, Biskup S, et al.: **Aberrant phase separation and nucleolar dysfunction in rare genetic diseases**. *Nature* 2023, **614**:564–571.

78. Mompeán M, Oroz J, Laurents D V.: **Do polyproline II helix associations modulate biomolecular condensates?** *FEBS Open Bio* 2021, **11**:2390–2399.

79. Morgan AA, Rubenstein E: **Proline: The Distribution, Frequency, Positioning, and Common Functional Roles of Proline and Polyproline Sequences in the Human Proteome**. *PLoS One* 2013, **8**:e53785.

80. Posey AE, Ruff KM, Harmon TS, Crick SL, Li A, Diamond MI, Pappu R V.: **Profilin reduces aggregation and phase separation of huntingtin N-terminal fragments by preferentially binding to soluble monomers and oligomers**. *J Biol Chem* 2018, **293**:3734–3746.

81. Krishnan K, Moens PDJ: **Structure and functions of profilins**. *Biophys Rev* 2009, **1**:71–81.

82. Choudhary A, Gandla D, Krow GR, Raines RT: **Nature of amide carbonyl-carbonyl interactions in proteins**. *J Am Chem Soc* 2009, **131**:7244–7246.

83. Wilhelm P, Lewandowski B, Trapp N, Wennemers H: **A crystal structure of an oligoproline PPII-Helix, at last**. *J Am Chem Soc* 2014, **136**:15829–15832.

84. Newberry RW, Raines RT: **The n→π* Interaction**. *Acc Chem Res* 2017, **50**:1838–1846.

85*. Urbanek A, Popovic M, Elena-Real CA, Morató A, Estaña A, Fournet A, Allemand F, Gil AM, Cativiela C, Cortés J, et al.: **Evidence of the Reduced Abundance of Proline cis Conformation in Protein Poly Proline Tracts**. *J Am Chem Soc* 2020, **142**:7976–7986.

Using SSIL samples and NMR, the authors study in a position-specific manner the *cis/trans* isomerization equilibrum of polyP tracts of different lengths in huntingtin. They show that the population of the *cis* isomer depends on the proline position and the length of the homorepeat. These features explain the stiffness of polyP.

86. Shen K, Calamini B, Fauerbach JA, Ma B, Shahmoradian SH, Serrano Lachapel IL, Chiu W, Lo DC, Frydman J: **Control of the structural landscape and neuronal proteotoxicity of mutant Huntingtin by domains flanking the polyQ tract.** *Elife* 2016, **5**:1–29.

87. Tartari M, Gissi C, Lo Sardo V, Zuccato C, Picardi E, Pesole G, Cattaneo E:

Phylogenetic comparison of huntingtin homologues reveals the appearance of a primitive polyQ in sea urchin. *Mol Biol Evol* 2008, **25**:330–338.

88. Zhang L, Kang H, Perez-Aguilar JM, Zhou R: **Possible Co-Evolution of Polyglutamine and Polyproline in Huntingtin Protein: Proline-Rich Domain as Transient Folding Chaperone**. *J Phys Chem Lett* 2022, **13**:6331–6341.

89. Dobitz S, Aronoff MR, Wennemers H: **Oligoprolines as Molecular Entities for Controlling Distance in Biological and Material Sciences**. *Acc Chem Res* 2017, **50**:2420–2428.

90. Schuler B, Lipman EA, Steinbach PJ, Kumkell M, Eaton WA: **Polyproline and the "spectroscopic ruler" revisited with single-molecule flourescence**. *Proc Natl Acad Sci U S A* 2005, **102**:2754–2759.

91. Lilliu E, Villeri V, Pelassa I, Cesano F, Scarano D, Fiumara F: **Polyserine repeats promote coiled coil-mediated fibril formation and length-dependent protein aggregation**. *J Struct Biol* 2018, **204**:572–584.

92. Xu S, Lai SK, Sim DY, Ang WSL, Li HY, Roca X: **SRRM2 organizes splicing condensates to regulate alternative splicing**. *Nucleic Acids Res* 2022, **50**:8599–8614.

93*. Lester E, Van Alstyne M, McCann KL, Reddy S, Cheng LY, Kuo J, Pratt J, Parker R: **Cytosolic condensates rich in polyserine define subcellular sites of tau aggregation**. *Proc Natl Acad Sci U S A* 2023, **120**:e2217759120.

In this study, the authors show that the polyS tracts present in SRRM2 and PNN mediate in a length-dependent manner the interaction of these proteins with Tau aggregates, found in some neurodegenerative diseases. Moreover, they show that polyS are necessary to create condensates that are preferential sites for Tau aggregate propagation.

94. Wolf A, Mantri M, Heim A, Müller U, Fichter E, Mackeen MM, Schermelleh L, Dadie G, Leonhardt H, Vénien-Bryan C, et al.: **The polyserine domain of the lysyl-5 hydroxylase Jmjd6 mediates subnuclear localization**. *Biochem J* 2013, **453**:357–370.

95. Rudich P, Watkins S, Lamitina T: **PolyQ-independent toxicity associated with novel translational products from CAG repeat expansions**. *PLoS One* 2020, **15**:e0227464.

96. Banez-Coronel M, Ranum LPW: **Repeat-associated non-AUG (RAN) translation: insights from pathology**. *Lab Investig* 2019, **99**:929–942.

97. Murrali MG, Piai A, Bermel W, Felli IC, Pierattelli R: **Proline Fingerprint in Intrinsically Disordered Proteins**. *ChemBioChem* 2018, **19**:1625–1629.

98. Felli IC, Pierattelli R: **13C Direct Detected NMR for Challenging Systems**. *Chem Rev* 2022, **122**:9468–9496.

99. Best RB, Zheng W, Mittal J: **Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association**. *J Chem Theory Comput* 2014, **10**:5113–5124.

100. Robustelli P, Piana S, Shaw DE: **Developing a molecular dynamics force field**

for both folded and disordered protein states. *Proc Natl Acad Sci U S A* 2018, **115**:E4758–E4766.

101. Song D, Luo R, Chen HF: **The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins**. *J Chem Inf Model* 2017, **57**:1166–1178.

102. Zhong B, Song G, Chen HF: **Balanced Force Field ff03CMAP Improving the Dynamics Conformation Sampling of Phosphorylation Site**. *Int J Mol Sci* 2022, **23**.

103*. Tang WS, Fawzi NL, Mittal J: **Refining All-Atom Protein Force Fields for Polar-Rich, Prion-like, Low-Complexity Intrinsically Disordered Proteins**. *J Phys Chem B* 2020, **124**:9505–9512.

Using experimental data, the authors have parametrised the popular ff99SBws force field in order to better describe the conformational features of polar residues Serine, Threonine and Glutamine. This modified force field (ff99SBws-STQ) promises more accurate structural description of low complexity, prion-like regions without compromising the results for the rest of the protein.

104. Fawzi NL, Parekh SH, Mittal J: **Biophysical studies of phase separation integrating experimental and computational methods**. *Curr Opin Struct Biol* 2021, **70**:78–86.

105. Tesei G, Schulze TK, Crehuet R, Lindorff-Larsen K: **Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties**. *Proc Natl Acad Sci U S A* 2021, **118**.

106. Rizuan A, Jovic N, Phan TM, Kim YC, Mittal J: **Developing Bonded Potentials for a Coarse-Grained Model of Intrinsically Disordered Proteins**. *J Chem Inf Model* 2022, **62**:4474–4485.

107. Regy RM, Thompson J, Kim YC, Mittal J: **Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins**. *Protein Sci* 2021, **30**:1371–1379.

108. Perdigão N, Heinrich J, Stolte C, Sabir KS, Buckley MJ, Tabor B, Signal B, Gloss BS, Hammang CJ, Rost B, et al.: **Unexpected features of the dark proteome**. *Proc Natl Acad Sci U S A* 2015, **112**:15898–15903.

109. Nielsen JT, Mulder FAA: **POTENCI: prediction of temperature, neighbor and pH-corrected chemical shifts for intrinsically disordered proteins.** *J Biomol NMR* 2018, **70**:141–165.

Figures:

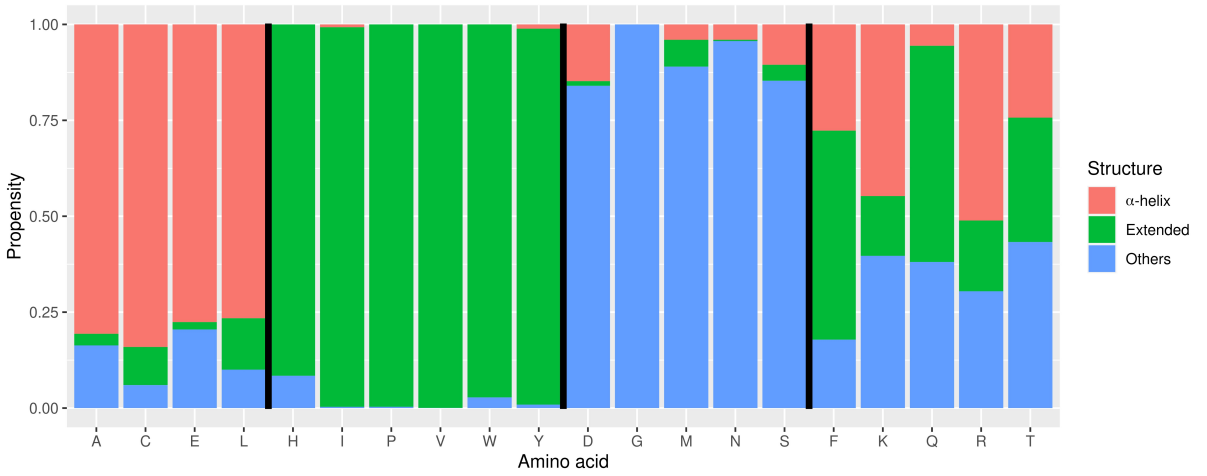

Figure 1. Secondary structure preferences in pure stretches of amino acids. Fraction of predicted α-helical, extended (β-strand and PP-II) and Others, which joins all other secondary structure combinations. Predictions were performed for 30-residue long polyX fragments with the LS2P program[39]. Note that only the central stretch was analysed. Amino acids are classified in four groups according to their preferred predicted conformation.

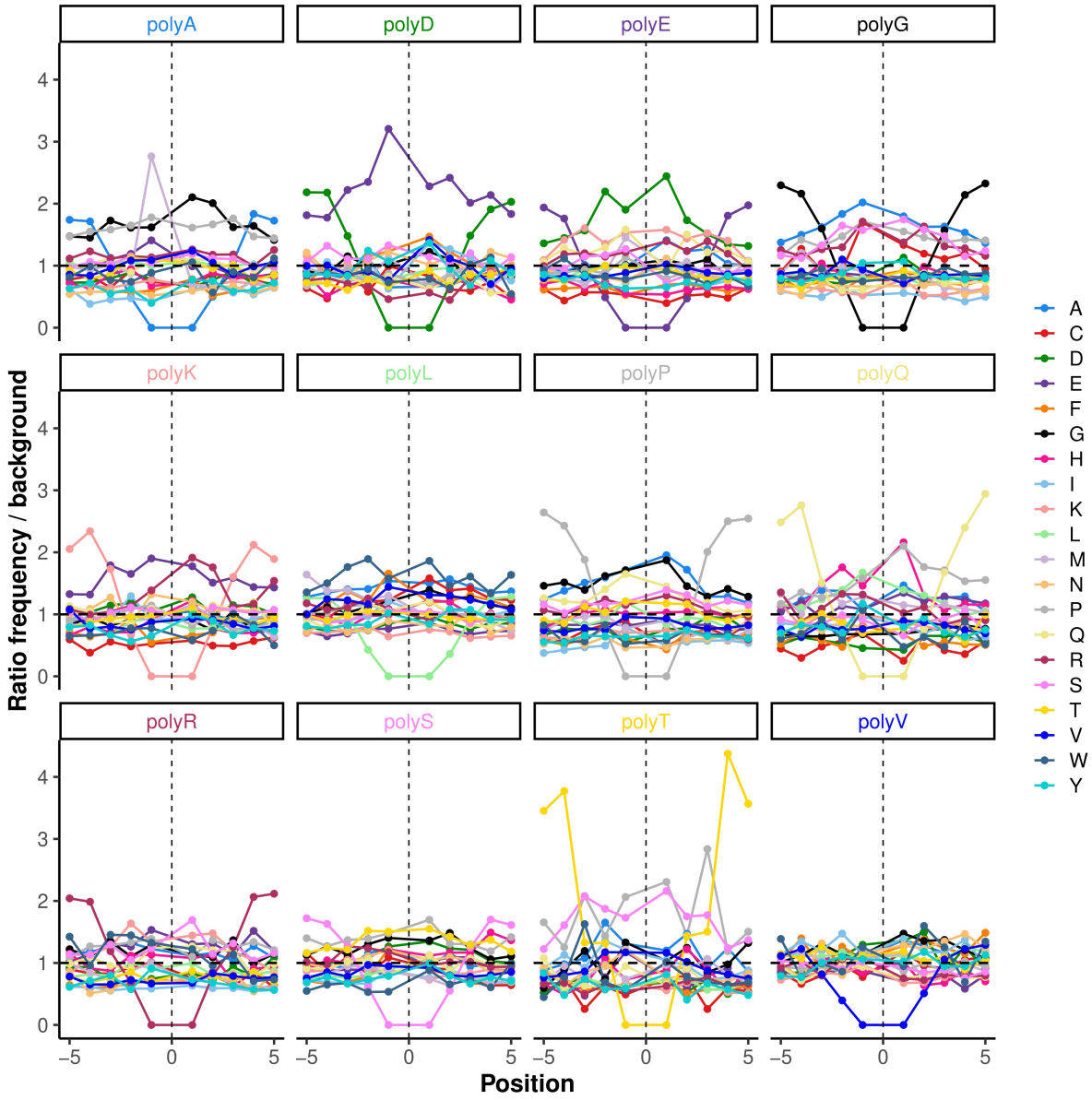

Figure 2. Ratio of amino acid frequency per position in relation to polyX regions (taken as position 0) versus the background amino acid prevalence in the human proteome. The complete human reference proteome from UniProtKB release v2023_01 (20,591 proteins) was used for the analysis. Only homorepeat types found at least 1000 times in the human proteome are shown.

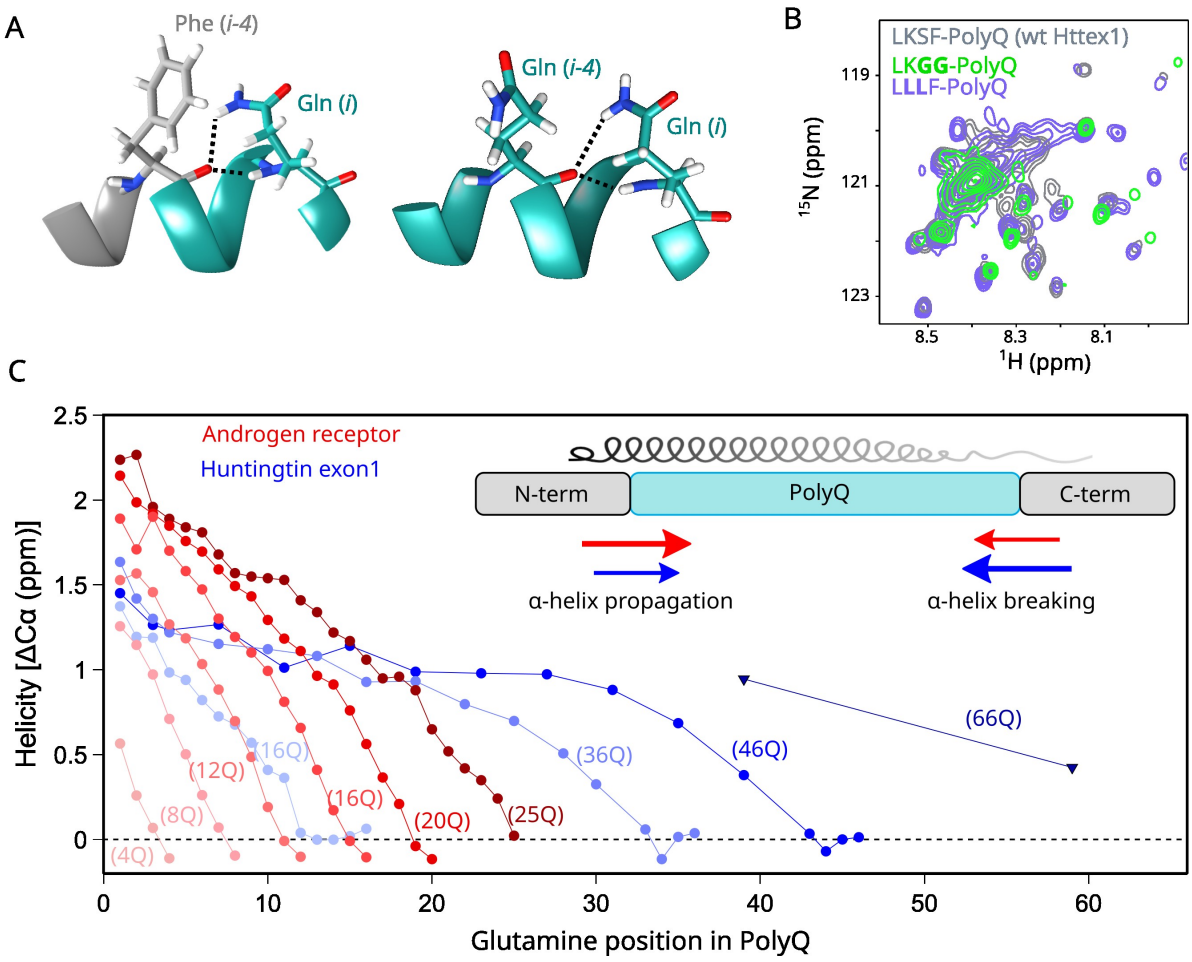

Figure 3. α-helix propagation along polyQ tracts. (A) Structural model of bifurcated hydrogen bonds between HttExon-1 N17 and polyQ (left), or between glutamines inside the homorepeat (right). (B) Zoom of overlapping ^{15}N-HSQC spectra of wt HttExon-1 with 46 glutamines, and two mutants that either reduce or increase the polyQ helical conformation (green and purple, respectively). (C) Cα secondary chemical shift (SCS) profile of AR (red) and HttExon-1 (blue) with polyQ tracts with different lengths. Experimental Cα chemical shifts are compared with a neighbour–corrected random coil database[109].