

Evolutionary Study of Protein Short Tandem Repeats in Protein Families

Pablo Mier *  and Miguel A. Andrade-Navarro 

Faculty of Biology, Institute of Organismic and Molecular Evolution, Johannes Gutenberg University Mainz, 55128 Mainz, Germany; andrade@uni-mainz.de

* Correspondence: munoz@uni-mainz.de

Abstract: Tandem repeats in proteins are patterns of residues repeated directly adjacent to each other. The evolution of these repeats can be assessed by using groups of homologous sequences, which can help pointing to events of unit duplication or deletion. High pressure in a protein family for variation of a given type of repeat might point to their function. Here, we propose the analysis of protein families to calculate protein short tandem repeats (pSTRs) in each protein sequence and assess their variability within the family in terms of number of units. To facilitate this analysis, we developed the pSTR tool, a method to analyze the evolution of protein short tandem repeats in a given protein family by pairwise comparisons between evolutionarily related protein sequences. We evaluated pSTR unit number variation in protein families of 12 complete metazoan proteomes. We hypothesize that families with more dynamic ensembles of repeats could reflect particular roles of these repeats in processes that require more adaptability.

Keywords: low complexity regions; repeats; protein sequence analysis

1. Introduction

Tandem repeats in proteins are periodic sequences consisting of several identical or similar units of residues adjacent to each other [1]. They are common in protein sequences [2,3]. The length of the units ranges from one to more than 60 [4]. Even if they emerge as the repetition of identical units, they may mutate and lead to degenerated tandem repeats. In the case of repeats forming assemblies of repeated structural units, even when degenerated, they will still conserve their general structure [5].

The detection of tandem repeats relies on a direct search of repetitive elements in a sequence or structure using either homology search [6], clustering [7], with Hidden Markov Models [8], or taking advantage of the modularity of protein structures as defined in PDB entries [9,10]. Computational approaches have been developed for the detection of tandem repeats in genomic DNA (e.g., TriCoLOR [11]) and proteins (e.g., REP2 [12]). RepeatsDB is a database of tandem repeats in protein structures [13], which is also used as a library to predict tandem repeats in RepeatsDB-lite [14].

The simplest form of tandem repeats are homorepeats, consisting of units of length one. Homorepeats such as polyglutamine and polyalanine regions have been reported to grow in length due to replication slippage [15,16]. This process occurs when the DNA polymerase encounters a repeated sequence and results in the addition of a new unit to the existing repeat or in the deletion of a repeated unit during the replication process [17].

In a recent study where we assessed simple repeats in protein sequences, we noted their high frequency and prevalence across protein families in many organisms [18]. In particular, we observed very often that some protein families display bursts of short repeats, usually in disordered regions, and we observed tendencies for particular repeat lengths in a few organisms. Particular protein families can have functions that put them under evolutionary pressure for tandem repeat unit variation because this variation gives them some selective advantage. For example, we observed the pressure to accumulate polyQ



Citation: Mier, P.; Andrade-Navarro, M.A. Evolutionary Study of Protein Short Tandem Repeats in Protein Families. *Biomolecules* **2023**, *13*, 1116. <https://doi.org/10.3390/biom13071116>

Academic Editor: Jeffrey Brender

Received: 25 May 2023

Revised: 6 July 2023

Accepted: 12 July 2023

Published: 13 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

in protein families with many protein interactions, and we used this as support for the function of polyQ in modulating protein interactions in vertebrates [19]. Similarly, we would expect that the characterization of the evolutionary expansion of given short repeats in particular protein families and species could reflect a functional role of such repeats. The association of regions biased for a few amino acids and functions are known in RG-rich regions in liquid–liquid phase separation and RNA binding-related proteins [20] and in RS-rich regions in splice-related plant proteins [21]. We hypothesized that a search for possible events of tandem duplication taking advantage of large numbers of sequences in protein families could serve to point to protein families and repeats under strong selective pressure for duplication, which could be helpful for their functional characterization.

Here, we present a strategy to identify variation in the number of units in protein short tandem repeats (pSTRs) in a set of homologous sequences using pairwise alignments. We provide this method as a web tool that takes as input a set of (ideally homologous) proteins and can be accessed at <http://cbdm-01.zdv.uni-mainz.de/~munoz/pSTR/> (accessed on 5 July 2023). Application of this method to protein families from 12 complete metazoan proteomes revealed the variable levels of evolution of these units across different protein families.

2. Materials and Methods

We downloaded a set of 70 expert-curated reference groups of orthologs (RefOGs) with proteins from 12 bilaterian species from the latest update of the OrthoBench benchmark suite [22]: *Caenorhabditis elegans* (cel), *Drosophila melanogaster* (dme), *Ciona intestinalis* (cin), *Danio rerio* (dre), *Tetraodon nigroviridis* (tni), *Gallus gallus* (gga), *Canis familiaris* (cfa), *Monodelphis domestica* (mdo), *Rattus norvegicus* (rna), *Mus musculus* (mmu), *Pan troglodytes* (ptr) and *Homo sapiens* (hsa). We also downloaded the complete reference proteomes of these species from UniProtKb release 2022_05 [23]. For the calculation of the pSTRs, we did not take into account protein sequences with ‘X’ residues.

3. Results

3.1. Protein Short Tandem Repeats Search Strategy

Protein Short Tandem Repeats (pSTRs) are residue units identical in sequence and directly adjacent in a protein sequence. We compute them in a sequence using the following procedure. For all lengths from two to half of the protein length (the theoretical maximum length of a pSTR to have more than one unit), a window of said length is slid through the protein sequence; if there are two or more consecutive equal units, the unit and the number of times it is repeated are stored (Figure 1). All patterns are reported, even if they are overlapping. We did not consider pSTRs of length one because it is relatively easy to find two consecutive identical amino acids in a sequence by chance. The “tandem” in pSTR refers to repetitions of the same unit; in Figure 1, for example, there are four different pSTRs, with units “ML”, “ACD”, “PQ” and “IL”.

When in a comparison between two protein sequences a pSTR unit fully aligns to a gap (green in Figure 1C) and is adjacent to a conserved pSTR (in pink), this highlights an event of gain or loss of a pSTR unit. To detect these, we do an all-versus-all pairwise comparison using MUSCLE v3.8.1551 with default parameters [24]. Two additional aligners were tested in the development phase: ClustalO v1.2.4 [25] and MAFFT v7.453 [26]; the results obtained with them were comparable, but MUSCLE was at least three times faster than the others. Our pipeline limits possible misalignments of the repeats by focusing only on pairwise alignments, and by using homologous proteins in the input dataset, so that the alignments are simplified by a high sequence similarity.

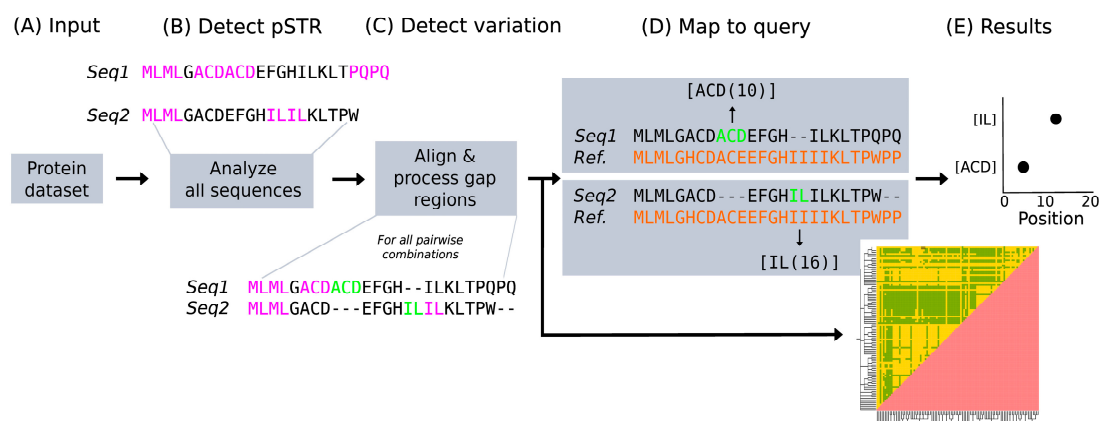


Figure 1. Workflow for the identification of pSTR unit number variation in protein datasets. (A) To identify pSTR and their unit number variation, we need a protein dataset in FASTA format. (B) Sequences are analyzed independently to search for pSTRs, in pink. (C) Then, sequences are aligned pairwise and unit variations are detected (in green) if they align to a gap and are repeated either N- or C-terminally. (D) Each sequence is then aligned with the query, which must be one of the proteins in the input dataset, to positionally map the pSTR with unit number variation and annotate them as [UNIT(mapped position)]. (E) Results are presented per pSTR with unit number variation and in a heatmap showing the protein pairs for which there is at least one of them.

For a direct comparison between the pSTR unit variants found within a family, they are mapped to positions in one sequence from the initial input dataset selected as query. Each protein for which we find pSTR variation is aligned with the query (with MUSCLE v3.8.1551 and default parameters [24]), and the position in the query sequence aligning with the central position of the extra unit is considered as its mapped position.

The query sequence itself may not share any residue with the extra unit since these are identified in alignments between different proteins within the input dataset, but the mapping to the query reflects the position of the repeat variant in the family (considering that the dataset comprises proteins from one protein family). In this way, it is possible to identify hotspots of repeat evolution of the entire family using the query sequence as reference.

3.2. Comparison of pSTRs in Metazoans

We identified the pSTRs in 12 complete metazoan proteomes. These are relatively frequent, with an average of 3.2% residues of the proteomes being in pSTRs, with frequency values ranging from 2.5% in *C. intestinalis* to 4.4% in *D. melanogaster* (Figure 2a). For this calculation, we took into account the proportion of residues in the proteomes participating in at least one pSTR, although we acknowledge that short pSTRs with units composed of very prevalent amino acids may be occurring randomly.

To add details to the differences in amino acid composition, we then calculated the amino acid frequencies in pSTRs comparing them to their individual amino acid background usage (Figure 2b). Amino acids with a low prevalence in proteomes tend to be even rarer in pSTRs and that is the case of cysteine, phenylalanine, histidine, methionine, tryptophan and tyrosine. The use of frequent amino acids in pSTRs is variable. Some amino acids are less used: leucine, isoleucine, asparagine, threonine, valine; some are more used: alanine, glutamic, glycine, proline, serine; and some are equally used: aspartic, lysine, glutamine, arginine.

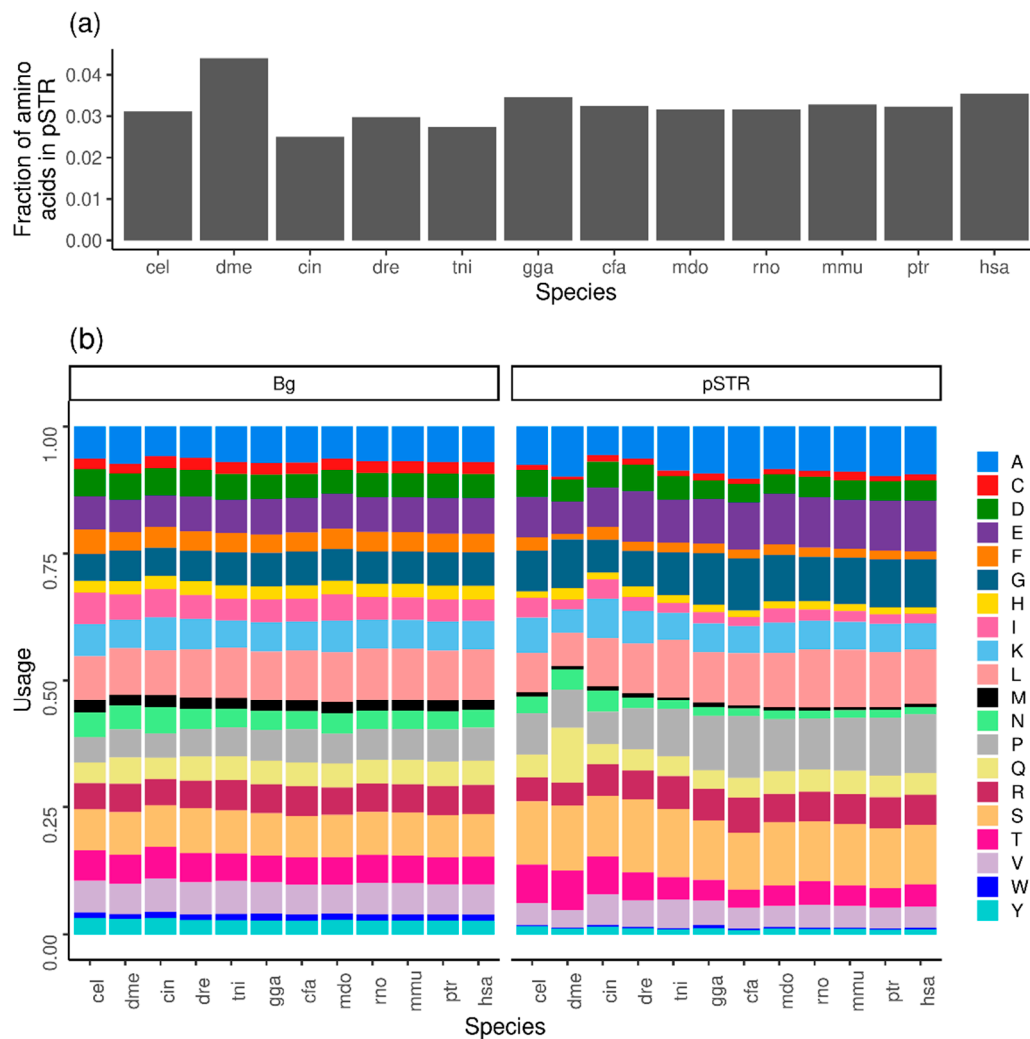


Figure 2. Properties of pSTRs in 12 species. **(a)** Fraction of amino acids in pSTRs. **(b)** Amino acid usage in background (whole proteome, **left**) and pSTRs (**right**). Species are annotated by a three-letter code (see Section 2).

Regarding species-specific trends, while the backgrounds are relatively constant, there is variation in the amino acid composition of pSTRs. Threonine and asparagine are more abundant in the pSTRs of *C. elegans*, *D. melanogaster* and *C. intestinalis*, and glutamine stands out in *D. melanogaster* (as already described [27]), to the best of our knowledge yet without a functional explanation. These results confirm and extend trends of repeat composition biases associated with particular species and suggest that different evolutionary pressures, that could be related to particular species and suggest that different evolutionary pressures, and that could be related to function, might influence the types and compositions of pSTRs.

3.3. Analysis of pSTR Unit Variation

Next, we downloaded the set of 70 Reference Orthologous Groups (RefOGs, or COGs) with proteins from the same 12 species as in the previous section, provided by OrthoBench (see Materials and Methods), the standard benchmark to assess the accuracy of orthogroup inference methods. For each COG, we calculated pSTR unit variation. For a total of 24,345 pSTRs, we identified 81 events of unit variation (Supplementary File S1). Only one COG had no pSTRs, indicating that while pSTRs take roughly 3% of proteomes, they are widely distributed across families. Differently, events of pSTR unit variation were identified in only 22 COGs (31%). We note that this number depends on the choice of species and that more events of pSTR unit variation might be found if more species are investigated.

Figure 3. Variation of unit number of two groups of tandem repeats in plant E3 ubiquitin ligase PQT3 and PTQ3L proteins. **(A)** Part of the multiple sequence alignment of *Arabidopsis thaliana* E3 ubiquitin ligase PQT3 (UniProtKB:F4JP52) and PQT3-like (UniProtKB:B9DFV2) proteins with homologs from *Capsella rubella*, *Raphanus sativus*, *Brassica oleracea*, *Tarenaya hassleriana* and *Herrania umbratica*. Perfect repeats (‘GP’ or ‘QPGFNGV’) or close variants are highlighted in pink and blue, respectively. **(B)** Phylogenetic tree derived from the alignment. Empty and full circles indicate events of duplication of the small (‘GP’) and large (‘QPGFNGV’) pSTRs, respectively. **(C)** AlphaFold model of the *A. thaliana* PQT3L protein structure (AF-B9DFV2-F1). Labels indicate amino acid positions and names of various sequence annotations given by UniProtKb, and of the two highlighted regions with variable pSTRs (in pink). Protein identifiers indicated in the alignment and tree (before species name) are from NCBI’s Entrez Protein.

The multiple sequence alignment of the *A. thaliana* PQT3L protein with homologs indicates the variation of repeats of ‘GP’ (one to four units), and close in sequence, the variation of seven amino acid repeats with sequence ‘QPGFNGV’ or similar (one to five units) (Figure 3A). There is no correlation between the number of units of each of the ensembles. The phylogenetic tree of these sequences can be used to infer consecutive and independent events of unit expansion (Figure 3B). We show the AlphaFold model of the *A. thaliana* protein structure just to illustrate that it leaves large regions not predicted, suggesting that they could be disordered (Figure 3C). This example suggests that PQT3L has a higher pressure for repeat unit variation than PQT3.

In the example, the mapping of unit variations obtained for the 10 sequences shown in the figure occupies six positions when mapped to the *A. thaliana* PQT3L: 508 and 509 for ‘GP’, 518 and 529 for ‘QPGFNGV’ and the similar ‘QPGFNGF’, and then there is unit ‘NNN’ at positions 651 and 654. In the MobiDB database of protein disorder [31], most of this protein is predicted to be disordered (74%) with the consensus of several disorder predicting methods. Considering that the disordered regions of *A. thaliana* PQT3L are predicted to be in amino acid positions 78–207, 236–241, 270–288 and 358–892, it is intriguing to note that most of the variable units occur in the small region shown in the alignment (508–529). Examples of protein-driven phase separation are known to occur in *A. thaliana* [32,33], and indeed E3 ubiquitin ligases are known to regulate phase separation and form part of phase-separated organelles [34]. Thus, it could be possible that this E3 plant family uses and regulates LLPS similarly.

3.4. Web Tool to Search for pSTR Unit Variation

We developed a web tool called pSTR (<http://cbdm-01.zdv.uni-mainz.de/~munoz/pSTR/>; accessed on 5 July 2023) that only needs as input a set of protein sequences in FASTA format. The tool identifies pSTR unit variation following the steps described above. The results are provided both raw (list of protein pairs with at least one event of pSTR unit variation, and their mapped position in the query) and processed. The latter includes a heatmap depicting the protein pairs that support at least one event of repeat unit variation (sequences in the same order as in the input file). In addition, the tool provides a representation of the positions of all variations mapped in the query.

Execution time depends on the number of sequences in the input file and on their length. As a guideline, a search with 100 sequences as input with a mean length of 226 amino acids executes in less than a minute.

As example for the use of the pSTR web tool, we provide the dataset containing the 10 homologs of PQT3/PQT3L used in Figure 3 (Supplementary File S2, and available in the web tool as an example dataset). The execution takes approximately 7 s.

Furthermore, we offer the user the option to run their analysis with the standalone version of the code. In it, there is the additional possibility to start the execution with just one protein in FASTA format. In that case, the program searches for homologs of the input protein (which is considered as the query sequence for the downstream mapping of the pSTR unit variants) in a set of 100 mammalian proteomes (Supplementary File S3), one protein per proteome (proteomes downloaded from the UniProtKB database release 2022_05 [23]). The search is performed with BLAST v2.9.0+ and default parameters [35]. The hypothetical selection of protein fragments as homologs would not compromise the results of the execution, but rather would minimize the number of pSTR with unit number variation obtained. Then, the search for pSTR starts with the generated multifasta file of all homologous sequences (or with the provided multifasta protein dataset).

4. Conclusions

We have developed a method to detect protein Short Tandem Repeats (pSTR) in protein sequences and to assess their unit variation with the aid of homologous sequences. Our analysis of a set of 70 curated groups of orthologs allowed us to report the properties of pSTRs and their variation. pSTRs cover 3% of sequences. While pSTRs are widely

distributed across protein families (only one had no pSTR), the 81 events of repeat unit variation we found occurred in just 22 groups, suggesting that these events have a tendency to cluster. We found a significant association of these events to families of proteins predicted to be involved in LLPS, which suggests that evolutionary pressure for repeat unit variation could come from an adaptive function related to protein phase formation.

Considering that our analysis of pSTR unit variation was based on a fixed set of 12 very taxonomically distant metazoan species, it is remarkable that we detected many events. Using these species was necessary for an automated procedure that allowed us to extract general conclusions about pSTR composition and unit variation at the level of Metazoa. A finer analysis is illustrated with our example of *A. thaliana* PQT3/PQT3L proteins. Manual exploration and iterative selection of homologs allowed to produce a refined set of plant proteins (more specifically, belonging to the malvids taxonomic clade), in which it is relatively easy to track individual events of pSTR unit duplication and mutation.

In any case, we understand that further work is needed to evaluate the association of particular types of pSTRs and mechanisms for their evolution with different taxa. In this first approach, we wanted to focus on the set of species featured in OrthoBench, to demonstrate the relevance of pSTR evolution as a driving force for variation in disordered regions of particular protein families.

To aid the exploration of particular families and the iterative selection of homologs, we have implemented our method as a web tool. With the continuous increase in the number of species and proteomes sequenced, we foresee that tracking the evolution of pSTRs will become increasingly simple. This will help the characterization of regions with repeats and low composition bias, revealing protein families and domains under evolutive pressure to change the number of repeat units, allowing us to understand how these regions evolve and to detect functions associated with them.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/biom13071116/s1>, Supplementary File S1: Results obtained per RefOG after searching for pSTR and unit number variation. For each RefOG, it includes an input multifasta file, the set of pSTR found pairwise between the sequences, the unique pSTR mapped to the query position in the sequence and their distribution and the set of pSTR found independently for all the sequences; Supplementary File S2: Set of homologs of proteins PQT3/PQT3L, available in the web tool pSTR as an example dataset; Supplementary File S3: Set of 100 mammalian proteomes available in the web tool pSTR to search for homologs of the protein of interest.

Author Contributions: P.M. conceived and carried out the project. M.A.A.-N. supervised the project. P.M. drafted the manuscript. All authors contributed in writing the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The authors received no specific funding for this publication.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: pSTR can be accessed from <http://cbdm-01.zdv.uni-mainz.de/~munoz/pSTR> (accessed on 5 July 2023), and a standalone version of the code downloaded from a dedicated GitHub repository <https://github.com/pmiemun/pSTR>.

Acknowledgments: We are grateful to our colleagues Dorothee Dormann (Biology, JGU), Andreas Wachter (Biology, JGU) and Katja Luck (Institute of Molecular Biology, Mainz) for fruitful discussions. This publication is based upon work from COST Action ML4NGP, CA21160, supported by COST (European Cooperation in Science and Technology).

Conflicts of Interest: The authors declare no conflict of interest.

References

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. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Marcotte, E.M.; Pellegrini, M.; Yeates, T.O.; Eisenberg, D. A census of protein repeats. *J. Mol. Biol.* **1999**, *293*, 151–160. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Jorda, J.; Kajava, A.V. Protein homorepeats sequences, structures, evolution, and functions. *Adv. Protein Chem. Struct. Biol.* **2010**, *79*, 59–88.
4. Kajava, A.V. Tandem repeats in proteins: From sequence to structure. *J. Struct. Biol.* **2012**, *179*, 279–288. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Andrade, M.A.; Perez-Iratxeta, C.; Ponting, C.P. Protein repeats: Structures, functions, and evolution. *J. Struct. Biol.* **2001**, *134*, 117–131. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Andrade, M.A.; Ponting, C.P.; Gibson, T.J.; Bork, P. Homology-based method for identification of protein repeats using statistical significance estimates. *J. Mol. Biol.* **2000**, *298*, 521–537. [\[CrossRef\]](#)
7. Jorda, J.; Kajava, A.V. T-REKS: Identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics* **2009**, *25*, 2632–2638. [\[CrossRef\]](#)
8. Biegert, A.; Söding, J. De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics* **2008**, *24*, 807–814. [\[CrossRef\]](#)
9. Walsh, I.; Sirocco, F.G.; Minervini, G.; Di Domenico, T.; Ferrari, C.; Tosatto, S.C.E. RAPHAEL: Recognition, periodicity and insertion assignment of solenoid protein structures. *Bioinformatics* **2012**, *28*, 3257–3264. [\[CrossRef\]](#)
10. Viet, P.D.; Roche, D.B.; Kajava, A.V. TAPPO: A combined method for the identification of tandem repeats in protein structures. *FEBS Lett.* **2015**, *589*, 2611–2619. [\[CrossRef\]](#)
11. Bolognini, D.; Magi, A.; Benes, V.; Korbel, J.O.; Rausch, T. TRiCoLoR: Tandem repeat profiling using whole-genome long-read sequencing data. *Gigascience* **2020**, *9*, giaa101. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Kamel, M.; Kastano, K.; Mier, P.; Andrade-Navarro, M.A. REP2: A Web Server to Detect Common Tandem Repeats in Protein Sequences. *J. Mol. Biol.* **2021**, *433*, 166895. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Paladin, L.; Bevilacqua, M.; Errigo, S.; Piovesan, D.; Mičetić, I.; Necci, M.; Monzon, A.M.; Fabre, M.L.; Lopez, J.L.; Nilsson, J.F.; et al. RepeatsDB in 2021: Improved data and extended classification for protein tandem repeat structures. *Nucleic Acids Res.* **2020**, *49*, D452–D457. [\[CrossRef\]](#)
14. Hirsh, L.; Paladin, L.; Piovesan, D.; Tosatto, S.C.E. RepeatsDB-lite: A web server for unit annotation of tandem repeat proteins. *Nucleic Acids Res.* **2018**, *46*, W402–W407. [\[CrossRef\]](#)
15. Albà, M.M.; Santibáñez-Koref, M.F.; Hancock, J.M. The comparative genomics of polyglutamine repeats: Extreme differences in the codon organization of repeat-encoding regions between mammals and *Drosophila*. *J. Mol. Biol.* **2001**, *52*, 249–259. [\[CrossRef\]](#)
16. Shoubridge, C.; Gecz, J. Polyalanine tract disorders and neurocognitive phenotypes. In *Madame Curie Bioscience Database*; Landes Bioscience: Austin, TX, USA, 2011. Available online: <https://www.ncbi.nlm.nih.gov/books/NBK51932/> (accessed on 5 July 2023).
17. Viguera, E.; Canceill, D.; Ehrlich, S. Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J.* **2001**, *20*, 2587–2595. [\[CrossRef\]](#)
18. Kamel, M.; Mier, P.; Tari, A.; Andrade-Navarro, M.A. Repeatability in protein sequences. *J. Struct. Biol.* **2019**, *208*, 86–91. [\[CrossRef\]](#)
19. Schaefer, M.H.; Wanker, E.E.; Andrade-Navarro, M.A. Evolution and function of CAG/polyglutamine repeats in protein–protein interaction networks. *Nucleic Acids Res.* **2012**, *40*, 4273–4287. [\[CrossRef\]](#)
20. Chong, P.A.; Vernon, R.M.; Forman-Kay, J.D. RGG/RG Motif Regions in RNA Binding and Phase Separation. *J. Mol. Biol.* **2018**, *430*, 4650–4665. [\[CrossRef\]](#)
21. Reddy, A.S.N.; Ali, G.S. Plant serine/arginine-rich proteins: Roles in precursor messenger RNA splicing, plant development, and stress responses. *Wiley Interdiscip. Rev. RNA* **2011**, *2*, 875–889. [\[CrossRef\]](#)
22. Emms, D.M.; Kelly, S. Benchmarking Orthogroup Inference Accuracy: Revisiting Orthobench. *Genome Biol. Evol.* **2020**, *12*, 2258–2266. [\[CrossRef\]](#) [\[PubMed\]](#)
23. UniProt Consortium. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [\[CrossRef\]](#)
25. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Mier, P.; Elena-Real, C.; Urbanek, A.; Bernadó, P.; Andrade-Navarro, M.A. 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. [\[CrossRef\]](#) [\[PubMed\]](#)

28. Kastano, K.; Mier, P.; Dosztányi, Z.; Promponas, V.J.; Andrade-Navarro, M.A. Functional Tuning of Intrinsically Disordered Regions in Human Proteins by Composition Bias. *Biomolecules* **2022**, *12*, 1486. [[CrossRef](#)]
29. Farahi, N.; Lazar, T.; Wodak, S.J.; Tompa, P.; Pancsa, R. Integration of Data from Liquid–Liquid Phase Separation Databases Highlights Concentration and Dosage Sensitivity of LLPS Drivers. *Int. J. Mol. Sci.* **2021**, *22*, 3017. [[CrossRef](#)]
30. Hardenberg, M.; Horvath, A.; Ambrus, V.; Fuxreiter, M.; Vendruscolo, M. Widespread occurrence of the droplet state of proteins in the human proteome. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 33254–33262. [[CrossRef](#)]
31. Piovesan, D.; Del Conte, A.; Clementel, D.; Monzon, A.M.; Bevilacqua, M.; Aspromonte, M.C.; Iserte, J.A.; Orti, F.E.; Marino-Buslje, C.; Tosatto, S.C.E. MobiDB: 10 years of intrinsically disordered proteins. *Nucleic Acids Res.* **2023**, *51*, D438–D444. [[CrossRef](#)]
32. Huang, S.; Zhu, S.; Kumar, P.; MacMicking, J.D. A phase-separated nuclear GBPL circuit controls immunity in plants. *Nature* **2021**, *594*, 424–429. [[CrossRef](#)]
33. Zhu, S.; Gu, J.; Yao, J.; Li, Y.; Zhang, Z.; Xia, W.; Wang, Z.; Gui, X.; Li, L.; Li, D.; et al. Liquid-liquid phase separation of RBGD2/4 is required for heat stress resistance in Arabidopsis. *Dev. Cell* **2022**, *57*, 583–597. [[CrossRef](#)]
34. Chin, L.-S.; Olzmann, J.A.; Li, L. Parkin-mediated ubiquitin signalling in aggresome formation and autophagy. *Biochem. Soc. Trans.* **2010**, *38*, 144–149. [[CrossRef](#)] [[PubMed](#)]
35. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.