Short Papers

Xsurvey: Web Tool to Query the Set of Homorepeats of all Reference Proteomes

Miguel A. Andrade-Navarro and Pablo Mier

***Abstract*—Homorepeats are low complexity regions in protein sequences composed of repetitions of one specific amino acid residue. There is currently no automatic way to compare the set of homorepeats between two or more species. Here we present Xsurvey, a web tool to query the set of homorepeats of 23,150 completely-sequenced proteomes. The polyX usage values can be easily compared visually, which simplifies the interpretation of the results.**

***Index Terms*—Homorepeats, protein sequence analysis, web tool.**

I. Introduction

Homorepeats, or polyX regions, are protein motifs in which one amino acid type is found many more times than expected by its amino acid usage. They can be formed by any amino acid, are present in all taxonomic branches of life, and evolve rapidly [1]. A few polyX types have been associated with functional roles, such as polyQ in mediating the assembly of large protein complexes, polyA in modulating protein function and stability, polyL as part of signal peptides, and polyG and polyS in aiding protein localization [2], [3], [4], [5], [6]. Other polyX types may not have yet an associated function, either because they do not have one, or because they have not been studied in detail.

PolyX regions are selected by applying a threshold on the number of identical residues (X) found in a sequence window (of length Y), represented as $>=$ X/Y. Commonly applied thresholds are $>=$ 4/6 (lax) and $>=$ 8/10 (stricter) [7]. There are two web tools available to search the set of homorepeats of a protein dataset or a proteome; however, they are limited to one proteome per execution (polyX2 [8]), or pre-computed for a small dataset of 122 proteomes (HraP [9]). Here we describe a simple web tool to query the set of polyX regions of all available completely-sequenced reference proteomes. The results are presented in a user-friendly format that requires no prior knowledge, making them accessible to users with limited bioinformatics experience.

II. Implementation

We downloaded all complete reference proteomes available from UniProtKB release 2022_01 [10], a total of 23150 proteomes, distributed as follows: 10561 viruses, 9601 bacteria, 372 archaea, 2616 eukaryotes. We pre-computed the set of polyX regions for all these proteomes using the standalone version of the polyX2 tool [8]; it took approximately 16 hours on a Lenovo Thinkpad 64-bit with 15.3 Gb of RAM and an Intel Core i7-8665U CPU @ 1.90 GHz $\times$ 8, running Ubuntu 22.04 LTS.

The available thresholds to detect homorepeats are $>=$ 4/6 (default) and $>=$ 8/10, implying a minimum number of identical residues in a region of the defined length. To start the execution of Xsurvey, the tool needs a set of species to look for their homorepeats. There are four ways to define this set:

A) Provide a list of taxonomic IDs (TaxIDs) separated by semicolon, i.e., "83333;9606", for Escherichia coli K12 and Homo sapiens.
B) By manual selection of species by name. The user must write at least three characters and then select a species from the drop-down list. The list is kept open until the user starts writing again in the text area.
C) Choose one pre-selected set of reference proteomes, from eukaryotes (20 species), fungi (10 species), chordates (15 species), mammals (15 species) or primates (14 species).
D) Let the tool select a set of 20 random species, from the total dataset of 23150 complete reference proteomes.

If the user starts the execution having provided information in more than one option, the order of preference is A > B > C > D, meaning that only the information with the higher preference will be used. If no information is provided, option D) is automatically selected. If we do not have information for a TaxID provided in option A (or the TaxID does not exist), a warning message will be shown. For options A and B, species are ordered on the x-axis of the figure in the order in which they are given. In option C there is a pre-defined order, and in option D the order is random. There is no limit to the number of input species in options A and B.

The output presents a table with the following information per species: TaxID, ProteomeID from UniProtKB, species name, taxonomical information, number of polyX regions, and a tab-separated file with the polyX regions. The polyX usage per species is then analyzed and plotted using the ggplot2 package v3.5.1 [11] (Fig. 1). Labels for species, in the format "Species name (TaxID)", are presented in the x-axis of the plot. The polyX usage per type is shown in the y-axis, with the values alphabetically ordered from top to bottom. The total number of polyX per species is placed on top of each species' column. The raw data used to generate the figure can be downloaded from the "Download" section above the aforementioned table.

III. Case Studies

Xsurvey can be used to obtain comparative insight into the usage of polyX regions. Here we illustrate this in a narrow and in a wide taxa, primates and eukaryotes, respectively.

Received 4 November 2024; revised 27 March 2025; accepted 31 March 2025. Date of publication 7 April 2025; date of current version 8 August 2025. This work was funded by the Beatriz Galindo BG23/00060 and in part by the Spanish Ministry of Science, Innovation and Universities. This publication is based upon work from COST Action ML4NGP under Grant CA21160, and in part by the COST (European Cooperation in Science and Technology). *(Corresponding author: Miguel A. Andrade-Navarro.)*

Miguel A. Andrade-Navarro is with the Institute of Organismic and Molecular Evolution, Faculty of Biology, Johannes Gutenberg University, 55128 Mainz, Germany (e-mail: andrade@uni-mainz.de).

Pablo Mier is with the Andalusian Centre for Developmental Biology, Faculty of Experimental Sciences, University Pablo de Olavide, 41013 Seville, Spain (e-mail: pmiemun@upo.es).

Xsurvey is freely available for public use at https://cbdm-01.zdv.uni-mainz.de/~munoz/xsurvey/.

Digital Object Identifier 10.1109/TCBBIO.2025.3557503

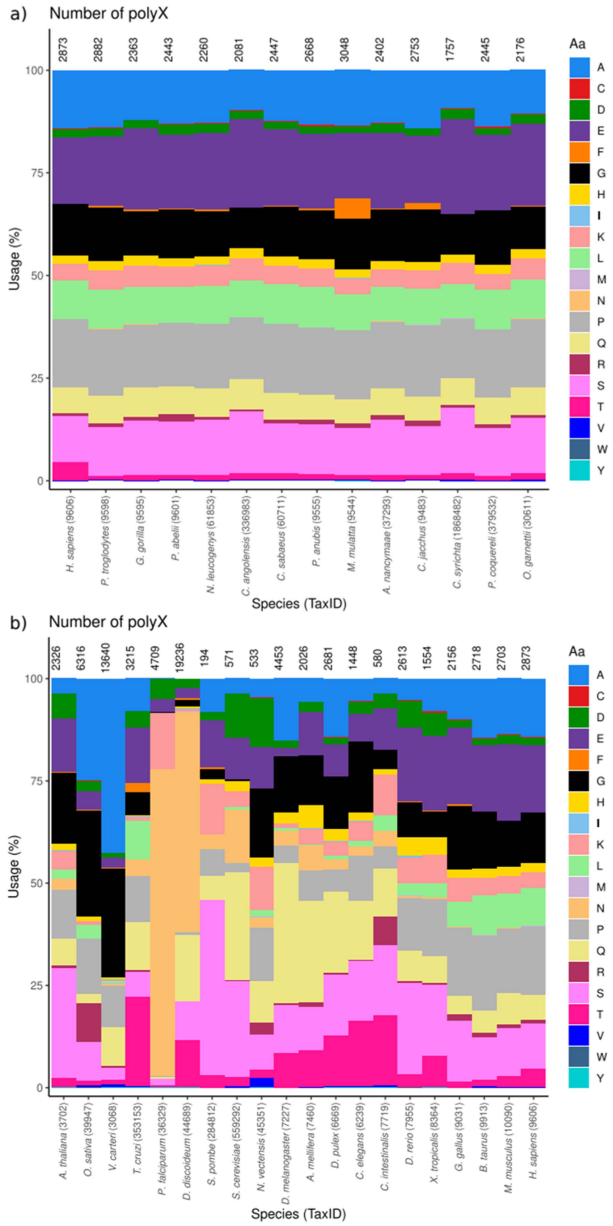

Fig. 1. Xsurvey representation of the polyX usage of a pre-selected set of proteomes from a) primates and b) eukaryotes, with threshold $> = 8/10$.

The total and relative abundance of most of the homorepeats is very similar amongst the pre-selected set of 14 primates (data not shown), with the default threshold of a minimum of four identical residues in a window of six residues (threshold $> = 4/6$). With this threshold, we include in the comparison short and long homorepeats. However, when we restrict the search to longer homorepeats (threshold $> = 8/10$), we see two major discrepancies (Fig. 1a). First, polyF in *Macaca mulatta* accounts for 5% of the total homorepeats in this species, but for 0% (human) to 1.7% (*Callithrix jacchus*) in the rest. A closer look to the 155 long polyF regions in *M. mulatta* shows that 49% of them overlap with transmembrane regions and 24% with signal peptides, using experimental and predicted positionally-located features in UniProtKB. The presence of these homorepeats in *M. mulatta* may be due to genome assembly errors. Second, polyT in *H. sapiens* accounts for 4.5% of the total homorepeats in the human proteome, a 3.4-fold enrichment compared to the rest of the primates. Most of these (101 out of the 129 polyT regions) are however due to one protein, mucin-2 (UniProtKB:Q02817), which presents a T-rich region that spans over 2300 residues. Without this protein, the polyT usage in human would be comparable to the polyT usage in the rest of the primates.

PolyF usage in most of the pre-selected set of eukaryotic proteomes is negligible, as we found in that of primates (Fig. 1b). On the other hand, polyT values are higher in invertebrates. It has been shown that polyT regions accelerate calcium carbonate formation [12]. This chemical compound is used by invertebrates to build their skeletons and shells [13]. To the best of our knowledge, there are no reports in literature describing a higher proportion of polyT regions in invertebrate species.

There is much more information that can be extracted from the same figure, for both known and unknown patterns. From the former, for example, the enrichment of polyN in *P. falciparum* and *D. discoideum*, and the depletion of polyI in eukaryotes [14]; from the latter, the enrichment of polyA and the depletion of polyS in *V. carteri*, compared to other plants. The comparison of the results obtained for all (threshold $> = 4/6$) or just for long (threshold $> = 8/10$) polyX regions is an additional level of complexity to be taken into account.

What has been discussed about Fig. 1 is only the tip of the iceberg; there are many more differences that can be observed, and that could be further examined by performing, for example, functional enrichments of the protein sets with each polyX type in each species. Based on the generated figure, it is up to the users to ask themselves these questions and query the results accordingly. The selection of species will decisively influence the conclusions that can be drawn from the figure that is generated.

References

[1] S. Chavali, A. K. Singh, B. Santhanam, and M. M. Babu, "Amino acid homorepeats in proteins," *Nat. Rev. Chem.*, vol. 4, pp. 420–434, 2020.
[2] K. Inoue and K. Keegstra, "A polyglycine stretch is necessary for proper targeting of the protein translocation channel precursor to the outer envelope membrane of chloroplasts," *Plant J.*, vol. 34, pp. 661–669, 2003.
[3] P. P. Labaj, G. G. Leparc, A. F. Bardet, G. Kreil, and D. P. Kreil, "Single amino acid repeats in signal peptides," *FEBS J.*, vol. 277, pp. 3147–3157, 2010.
[4] M. H. Schaefer, E. E. Wanker, and M. A. Andrade-Navarro, "Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks," *Nucleic Acids Res.*, vol. 40, pp. 4273–4287, 2012.
[5] A. Wolf et al., "The polyserine domain of the lysyl-5 hydroxylase Jmjd6 mediates subnuclear localization," *Biochem. J.*, vol. 453, pp. 357–370, 2013.
[6] I. Pelassa, D. Corá, F. Cesano, F. J. Monje, P. G. Montarolo, and F. Fiumara, "Association of polyalanine and polyglutamine coiled coils mediates expansion disease-related protein aggregation and dysfunction," *Hum. Mol. Genet.*, vol. 23, pp. 3402–3420, 2014.
[7] P. Mier, C. Elena-Real, A. Urbanek, P. Bernadó, and M. A. Andrade-Navarro, "The importance of definitions in the study of polyQ regions: A tale of thresholds, impurities and sequence context," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 306–313, 2020.
[8] P. Mier and M. A. Andrade-Navarro, "PolyX2: Fast detection of homorepeats in large protein datasets," *Genes*, vol. 13, 2022, Art. no. 758.
[9] M. Y. Lobanov, I. V. Sokolovskiy, and O. V. Galzitskaya, "HRaP: Database of occurrence of HomoRepeats and patterns in proteomes," *Nucleic Acids Res.*, vol. 42, pp. D273–D278, 2014.
[10] The UniProt Consortium, "UniProt: The Universal protein knowledgebase in 2023," *Nucleic Acid Res.*, vol. 51, pp. D523–D531, 2023.
[11] H. Wickham, *Ggplot2: Elegant Graphics For Data Analysis*. Berlin, Germany: Springer, 2016, ISBN 978-3-319-24277-4.
[12] C. Liu, W. Zhang, and H. Liu, "Threonine and polythreonine accelerate calcium carbonate formation," *Cryst. Growth Des.*, vol. 24, pp. 892–898, 2024.
[13] H. A. Lowenstam and S. Weiner, *On Biomineralization*. London, U.K.: Oxford Univ. Press, 1989, ISBN 978-0-195-04977-0.
[14] P. Mier, G. Alanis-Lobato, and M. A. Andrade-Navarro, "Context characterization of amino acid homorepeats using evolution, position, and order," *Proteins*, vol. 85, pp. 709–719, 2017.