Computational and Structural Biotechnology Journal 21 (2023) 5408–5412

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Short Communication

The nucleotide landscape of polyXY regions

Pablo Mier [*], Miguel A. Andrade-Navarro

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

ARTICLE INFO

Keywords:
Protein sequence analysis
Low complexity regions
Linear motifs
PolyXY

ABSTRACT

PolyXY regions are compositionally biased regions composed of two different amino acids. They are classified according to the arrangement of the two amino acid types ‘X′ and ‘Y′ into direpeats (composed of alternating amino acids, e.g. ‘XYXYXY’), joined (composed of two consecutive stretches of each amino acid, e.g. ‘XXXYYY’) and shuffled (other arrangements, e.g., ‘XYXXYY’). They have been characterized at the amino acid level in all domains of life, and are described as often found within intrinsically disordered regions. Since DNA replication slippage has been proposed as a driver of repeat variation, and given that some polyXY have a repetitive nature, we hypothesized that characterizing the nucleotide coding of various types of polyXY could give hints about their origin and evolution. To test this, we obtained all polyXY regions in the human transcriptome, categorized them, and studied their coding nucleotide sequences. We observed that polyXY exacerbates the codon biases, and that the similarity between the X and Y codons is higher than in the background proteome. Our results support a general mechanism of emergence and evolution of polyXY from single-codon polyX. PolyXY are revealed as hotspots for replication slippage, particularly those composed of repeats: joined and direpeat polyXY. Interconversion to shuffled polyXY disrupts nucleotide repeats and restricts further evolution by replication slippage, a mechanism that we previously observed in polyX. Our results shed light on polyXY composition and should simplify the determination of their functions.

1. Introduction

Protein sequences may contain regions that differ in composition from the normal, random distribution of amino acids by having a reduced number of amino acid types. These regions are called low complexity regions (LCRs). This is a relaxed definition that implies a difference between a region and a protein dataset serving as a background (a proteome, for example) in amino acid frequency, repetitiveness, or both [1]. The simplest LCR, called homorepeat, is composed of only one type of amino acid and defined as a tract of consecutive repeated amino acids. Given ‘X′ as a repeating residue, they are also called polyX regions. They are abundant in eukaryotes (comprising ~15% of eukaryotic proteomes), and can be made up of almost any amino acid [2]. Different types of polyX regions have been associated with diverse functions: polyQ in mediating protein-protein interactions [3] and transcriptional activation [4], polyL, polyG, polyH and polyS in aiding protein localization [2,5–7], and polyA in transcriptional repression [8], among others [9].

Another simple LCR are polyXY, regions composed of two different types of amino acids (X and Y). The arrangement of the two different amino acids is taken into account to classify these regions into three categories: direpeats (e.g., ‘XYXYXY’), joined (e.g., ‘XXXYYY’) and shuffled (e.g., ‘XYYYXXY’).

So far, there has only been a comprehensive general characterization of polyXY regions in terms of taxonomic abundance [10] and structural properties [11]. They are more abundant than polyX regions and usually overlap with intrinsically disordered regions, although they can induce helical (polyEK, polyER) or beta (polyGS, polyEP) conformations [11, 12]. On the other hand, in recent years studies have been published describing the functional features of specific XY pairs, mostly in the form of XY-rich regions: polyDE is related to chromatin metabolism [13]; R/G-rich regions are involved in transcription, splicing and mRNA translation [14]; the polyGS of the human SMN protein (UniProt: Q16637) is involved in the interaction of SMN with proteins Gemin2 and Gemin8 [15].

An important question that remains under investigation is the origin of polyXY regions. Replication slippage has been well described to shape the evolution of polyX, evidenced by codon biases in some homorepeats [16] and by the existence of genetic diseases that result from codon expansion [17]. The widespread existence of repeats within intrinsically

[*] Corresponding author.
E-mail address: munoz@uni-mainz.de (P. Mier).

https://doi.org/10.1016/j.csbj.2023.10.054
Received 2 October 2023; Received in revised form 30 October 2023; Accepted 30 October 2023
Available online 31 October 2023
2001-0370/© 2023 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

disordered regions [18], which emerge abundantly even in some giant virus lineages [19], hints at the possibility that polyXY could also result from replication slippage, particularly when they are repetitive in nature. We hypothesize that if this were the case, there should be traces of this origin at the nucleotide level. To test this, we studied the polyXY regions in the human transcriptome, paying particular attention to the degree of synonymous codon usage in different polyXY categories.

2. Methods

We downloaded the human transcriptome from Ensembl (BioMart GRCh38p13 dataset) [20], limiting it to protein-coding transcripts, for a total of 98,711 transcripts. We filtered this dataset to keep only one random transcript per gene (22,518 sequences). Then, we used the standalone version of the XYs tool [10], with default parameters, to search for polyXY regions in the proteins encoded in these sequences. These parameters are: two different amino acids are required in a window of 6 amino acids, with a minimum of two occurrences per amino acid. Once a polyXY is located with these parameters, the window is extended until a 6-amino acid window that does not match is found. The polyXY regions were categorized based on the arrangement of the amino acids in direpeats (the two different amino acids alternate, e.g., XYXYXY, also allowing one unpaired X or Y, which can be at the termini of the region – e.g., XYXYXY**X** – or between direpeats – e.g., XY**Y**XYXY), joined (a stretch of one amino acid is followed by a stretch of the other amino acid, e.g., XXXYYYY), and shuffled (other order of amino acids, e.g., XXYXYYX).

We calculated similarity between codons using the Hamming distance [21], counting the number of mismatched nucleotides at analogous positions of two codons.

3. Results and discussion

3.1. Analysis of the polyXY regions in the human transcriptome

We detected polyXY regions in the human transcriptome using a published approach with default parameters [10] (see Methods). We identified 20,983 polyXY regions in a dataset of 22,518 human transcripts; 9952 transcripts had at least one polyXY (that is, 44% of human proteins; Supplementary File 1).

Joined polyXY (composed of a polyX followed by a polyY) account for 13.5% of cases (2824 regions covering 0.15% of the transcriptome). Direpeat polyXY (formed by alternating X and Y, allowing for one unpaired X or Y, see Methods) are 7.3% of the polyXY regions (1527 regions covering 0.09% of the transcriptome). All other polyXY, categorized as shuffled, represent 79.2% of the regions (16,632 polyXY covering 0.96% of the transcriptome). PolyXY composed of six codons are the most prevalent (13,680 of the 20,983 regions, 65%), with the regions classified as shuffled being the most abundant (Supplementary File 2). The longest polyXY is a polyAQ direpeat of 59 amino acids in the transcription elongation regulator 1 (Ensembl: ENST00000296702).

Together, all polyXY found cover 1.20% of the human transcriptome (similar to the 1.18% previously reported for Eukaryotes [10]).

The codon usage in polyXY regions is very similar to that of the

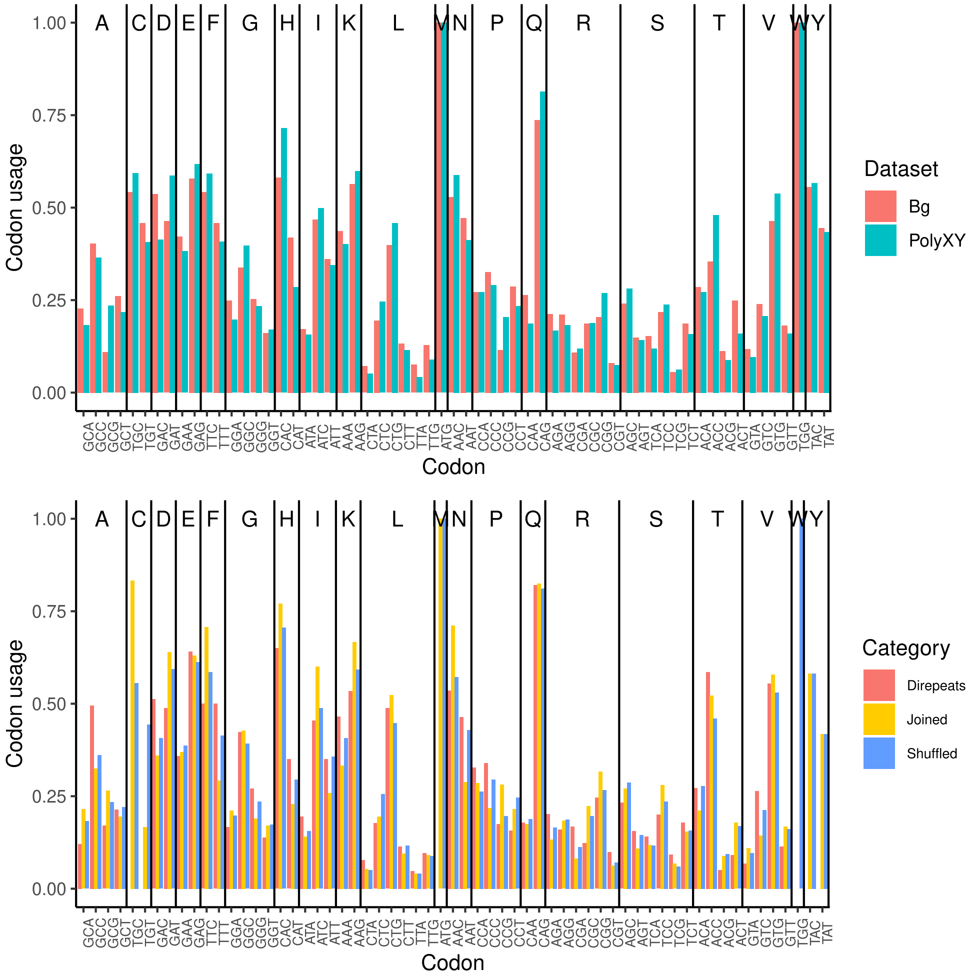

Fig. 1. Codon usage in polyXY regions. A) Codon usage in polyXY compared to the background of the complete proteome. B) Codon usage per polyXY category. Data is shown for codons occurring 10 or more times.

background (Fig. 1A). For fifteen amino acids, the most frequent codon in the background is even more frequent in the polyXY: for cysteine, glutamic acid, phenylalanine, glycine, histidine, isoleucine, lysine, leucine, asparagine, glutamine, serine, threonine, valine and tyrosine. Two, methionine and tryptophan, are encoded by only one amino acid. Only for three, alanine, aspartic acid and arginine, the most frequent codon in the background is less frequent in the polyXY. This indicates that polyXY tends to exacerbate codon usage bias. By polyXY category, the highest usage of a codon for each amino acid occurs within joined polyXY or direpeats polyXY for the majority of amino acids, twelve and four, respectively (Fig. 1B). Joined and direpeats polyXY are formed by amino acid repeats and can correspond to repeated codons. This supports that replication slippage, which has been proposed to be responsible for length variation in regions with homorepeats [22,23] and short tandem repeats [24], could also explain the formation of joined polyXY and direpeats polyXY.

3.2. *Joined polyXY are enriched in regions consisting of only two codon types*

If replication slippage were involved in originating joined polyXY and direpeats polyXY, we hypothesized that many of those polyXY should be encoded by a single codon type for amino acids X and Y (two-codon polyXY). In our dataset, two-codon polyXY are rare (6.1%). Consistent with our hypothesis, joined polyXY are more frequent among two-codon polyXY (22.5%) than in the complete set of polyXY (13.4%; Table 1). Differently, direpeat polyXY are only modestly enriched in two-codon polyXY: 8.9% versus 7.3% of direpeat polyXY among all polyXY. We interpret this result as an indication that there may be a lower selection pressure to keep identical codons in direpeat polyXY compared to joined polyXY.

3.3. *The X-codon and Y-codon in two-codon polyXY are significantly similar*

We hypothesized that if two-codon polyXY were frequently encoded by an X-codon and a Y-codon with only one nucleotide difference, this could explain the emergence of polyXY from mutations of a single-codon polyX.

To investigate the similarity between codons in two-codon polyXY, we calculated the Hamming distance (Hd) between codons in the top 10 most prevalent two-codon polyXY (Table 2; Supplementary File 3). As background, we computed the mismatches between codon combinations coding for each of the two amino acids forming the polyXY, chosen according to codon frequency. For some amino acid pairs, the result is always 1 because all pairs of codons are at distance 1 (e.g., for D/E, the two codons encoding each amino acid start with “GA”). For other pairs, the minimum possible distance is 2 (e.g., for E/L, the codons for glutamic acid start with “GA” and the codons for Leucine start with “CT” or with “TT”). Thus, when evaluating the distance in two-codon polyXY and in the background, it is necessary to account for the minimum and maximum possible distances, considering all possible codon comparisons.

The most interesting result is that, when variable results are possible, for the 10 most prevalent two-codon polyXY, the similarity between the codons is greater than in a background considering codon usage. This is indicative that polyXY tend to arise from mutations of a polyX. This tendency is stronger for some polyXY. For example, for polyGS, the mean distance in two-codon polyGS is 1.32, whereas in the background is 2.35 (Table 2). Out of 35 two-codon polyGS, 13 are joined, which is a high frequency, and 2 are direpeats. 32 of 35 encode Glycine with GGC (which is the most frequent codon) and 27 of 35 encode Serine with AGC (which is also the most frequent).

Table 1
PolyXY encoded by two codon types.

	Two-codon polyXY		All polyXY	
	Number	Frequency	Number	Frequency
Direpeats	113	8.9%	1527	7.3%
Joined	287	22.5%	2824	13.4%
Shuffled	875	68.6%	16,632	79.3%
Total	1275		20,983	

Table 2
Top 10 most prevalent types of two-codon polyXY regions. Min and Max indicate the minimum and maximum possible Hamming distance between two codons for the corresponding X and Y amino acids.

Pair	#polyXY	# of two-codon polyXY	Mean codon X to codon Y distance in two-codon polyXY	Min	Background based on codon usage	Max
D/E	830	139	1.00	1.00	1.00	1.00
E/K	505	63	1.22	1.00	1.49	2.00
E/L	393	54	2.07	2.00	2.63	3.00
E/Q	148	36	1.02	1.00	1.46	2.00
P/Q	389	34	1.79	1.00	1.84	2.00
G/S	875	35	1.32	1.00	2.35	3.00
L/Q	247	24	1.40	1.00	1.78	3.00
A/E	482	46	1.62	1.00	1.84	2.00
A/L	745	41	2.36	2.00	2.80	3.00
L/P	640	38	1.45	1.00	2.00	3.00

In a similar fashion, for polyEL we find a mean distance of 2.07, very close to the possible minimum of 2, while the background is 2.63. In this case, only 4 of 54 polyEL are joined, and just 8 are direpeats. Again, the codons used are the ones that are more frequent for each amino acid: GAG for Glutamic acid and CTG for Leucine. For polyEQ, the mean is 1.02 for a background of 1.46, and 7 of 36 polyEQ are joined with just two direpeats. Again, the codons used are the most frequent: 34 use GAG for Glutamic acid, and 33 use CAG for Glutamine.

There are cases where the similarity is not so different from the background. As an example, we present polyAE, with a mean distance of 1.62 compared to 1.84 in the background. 13 of 46 are joined and 2 are direpeats. In this case, the bias to use one codon is strong for Glutamic acid (40 use GAG, the most frequent codon). On the contrary, for Alanine only 16 use the most frequent codon GCC, another 16 use GCG, 9 use GCT, and 5 use GCA (the least frequent). Only 18 of the 46 polyAE have codons at distance 1. Because the Glutamic acid most frequent codon is GAG, the only codon of Alanine at distance 1 is GCG, which is the least frequent, with the most frequent codon GCC situated at distance 2.

Taken together, our results support the origin of polyXY from the mutation of a polyX, which was expanded by replication slippage of a frequent codon, followed by mutation into another frequent codon, followed by further replication slippage. This favors particular X and Y combinations.

3.4. *The codon similarity within the polyXY depends on the category*

To further investigate whether replication slippage might be more involved in the evolution of polyXY with amino acid repeats, joined and direpeats, than shuffled polyXY, we studied the distance between codons of particular amino acids in the three types of polyXY.

For this, we chose a random pair of codons for amino acid X and for amino acid Y for each polyXY and computed the average value in each category. As the background, we chose a similar number of codon pairs for each amino acid, distributed according to codon usage (Fig. 2).

Methionine and Tryptophan are encoded by a single codon. For 16 of the other 18 amino acids, the distance in polyXY was smaller than in the background, except for Phenylalanine and Isoleucine, for which shuffled

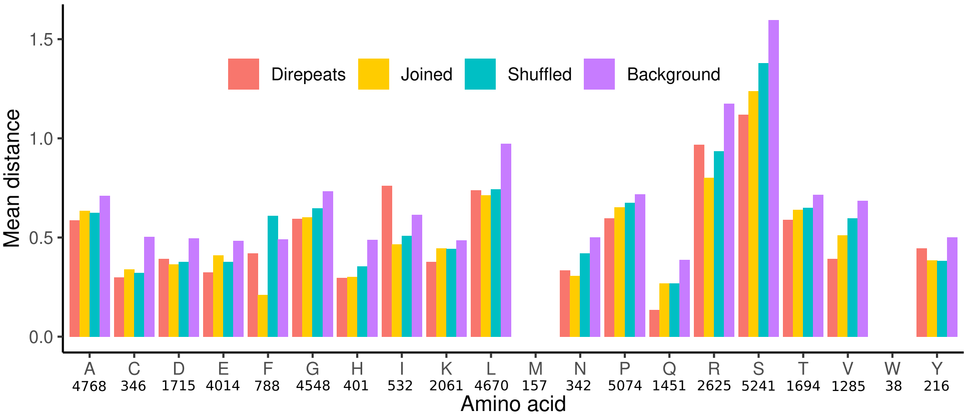

Fig. 2. Mean Hamming distance between codons for an amino acid in polyXY. Codons were sampled from corresponding polyXY (one pair of codons for X and for Y) or from a background using codon usage. The number under each amino acid label indicates the number of polyXY cases analysed.

and direpeats had larger distances than the background, respectively. Comparing the three polyXY types, polyXY with repeats were the ones with the shortest distance for most of the amino acids (11 and 6 times, for direpeats and joined, respectively), whereas shuffled polyXY were the ones with the shortest distance only for one amino acid (Tyrosine), and this by a small margin. These results add support to the participation of replication slippage in the evolution of polyXY with repeats.

4. Conclusions

We have characterized the codon usage of polyXY regions in the human transcriptome. Replication slippage was previously shown to be involved in the evolution of polyX [22,23]. Our results support that polyXY emerge from mutations of polyX regions of one codon, and that replication slippage is responsible for its subsequent evolution. This has two consequences. On the one hand, since Y originates from mutations of X, this imposes certain biases on the types of amino acids that become paired in a polyXY. On the other hand, very frequent codons become even more frequent within polyXY regions as these become hotspots for codon duplication. In the case of polyXY with repeats (direpeats and joined), replication slippage plays an important role.

From our studies of polyX [25] and short tandem repeats [24], we understand that repeats at the amino acid level provide rapidly evolving properties and adaptability to protein sequences and might be advantageous in some functional contexts, and that reducing nucleotide repeats using synonymous codons can be used to "freeze" this evolutionary variability. It is possible that the emergence of shuffled polyXY corresponds to this "freezing" process, which could be facilitated by an interconversion of X to Y, when they are encoded by codons that are both frequent and within Hamming distance 1. In this respect, shuffled polyXY could originate from mutations of repeated polyXY that keep the local composition bias, while eliminating codon repeats and potential replication slippage. Our study has expanded the role of replication slippage in the generation of low complexity regions and sheds light on the particularities of codon usage within these regions. While our results support a general mechanism for polyXY evolution, future work should explore the particular biases observed regarding certain amino acids that deviate from this mechanism and might have functional implications. Such studies should be extended to organisms other than humans.

Funding

This work was supported with funds from H2020-MSCA-RISE project REFRACT - GA No. 823886.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Data Availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Acknowledgements

This publication is based upon work from COST Action ML4NGP, CA21160, supported by COST (European Cooperation in Science and Technology).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.10.054.

References

[1] Mier P, Paladin L, Tamana S, Petrosian S, Hajdu-Soltesz B, et al. Disentangling the complexity of low complexity proteins. Brief Bioinform 2020;21:458–72.
[2] Mier P, Alanis-Lobato G, Andrade-Navarro MA. Context characterization of amino acid homorepeats using evolution, position, and order. Proteins 2017;85:709–19.
[3] Schaefer MH, Wanker EE, Andrade-Navarro MA. Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks. Nucleic Acids Res 2012;40:4273–87.
[4] Gerber HP, Seipel K, Georgiev O, Höfferer M, Hug M, et al. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 1994;263:808–11.
[5] Inoue K, Keegstra K. A polyglycine stretch is necessary for proper targeting of the protein translocation channel precursor to the outer envelope membrane of chloroplasts. Plant J 2003;34:661–9.
[6] Salichs E, Ledda A, Mularoni L, Albà MM, de la Luna S. Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. PLoS Genet 2009;5:e1000397.
[7] Wolf A, Mantri M, Heim A, Müller U, Fichter E, et al. The polyserine domain of the lysyl-5 hydroxylase Jmjd6 mediates subnuclear localization. Biochem J 2013;453:357–70.
[8] Galant R, Carroll SB. Evolution of a transcriptional repression domain in an insect Hox protein. Nature 2002;415:910–3.
[9] Chavali S, Singh AK, Santhanam B, Babu MM. Amino acid homorepeats in proteins. Nat Rev Chem 2020;4:420–34.
[10] Mier P, Andrade-Navarro MA. Regions with two amino acids in protein sequences: a step forward from homorepeats into the low complexity landscape. Comput Struct Biotechnol J 2022;20:5516–23.
[11] Gonçalves-Kulik M, Mier P, Kastano K, Cortés J, Bernadó P, et al. Low complexity induces structure in protein regions predicted as intrinsically disordered. Biomolecules 2022;12:1098.
[12] Gonçalves-Kulik M, Schmid F, Andrade-Navarro MA. One step closer to the understanding of the relationship IDR-LCR-structure. Genes 2023;14:1711.

[13] Shukla S, Lazarchuk P, Pavlova MN, Sidorova JM. Genome-wide survey of D/E repeats in human proteins uncovers their instability and aids in identifying their role in the chromain regulator ATAD2. iScience 2022;25:105464.

[14] Chong PA, Vernon RM, Forman-Kay JD. RGG/RG motif regions in RNA binding and phase separation. J Mol Biol 2018;430:4650–65.

[15] Wu C-Y, Curtis A, Choi YS, Maeda M, Xu MJ, et al. Identification of the phosphorylation sites in the survival motor neuron protein by protein kinase A. Biochim Biophys Acta 2011;1814:1134–9.

[16] Mier P, Andrade-Navarro MA. Between interactions and aggregates: the PolyQ balance. Genome Biol Evol 2021;13:evab246.

[17] Hancock JM. Simple sequences and the expanding genome. Bioessays 1996;18:421–5.

[18] Kamel M, Mier P, Tari A, Andrade-Navarro MA. Repeatability in protein sequences. J Struct Biol 2019;208:86–91.

[19] 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.

[20] Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2022. Nucleic Acids Res 2002;50:D988–95.

[21] Waggener B. (1995) Pulse Code Modulation Techniques. Springer p. 206. ISBN: 9780442014360.

[22] Albà MM, Santibáñez-Koref MF, Hancock JM. The comparative genomics of polyglutamine repeats: extreme differences in the codon organization of repeat-encoding regions between mammals and Drosophila. J Mol Evol 2001;52:249–59.

[23] Shoubridge C, Gecz J. Polyalanine tract disorders and neurocogniive phynotypes. Madame Curie Bioscience Database. Austin (TX). Landes Bioscience; 2011 (Available from), ⟨https://www.ncbi.nlm.nih.gov/books/NBK51932/⟩.

[24] Mier P, Andrade-Navarro MA. Evolutionary study of protein short tandem repeats in protein families. Biomolecules 2023;13:1116.

[25] Mier P, Andrade-Navarro MA. Glutamine codon usage and polyQ evolution in primates depend on the Q stretch length. Genome Biol Evol 2018;10:816–25.