Journal of Structural Biology 215 (2023) 108023

Contents lists available at ScienceDirect

Journal of Structural Biology

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

A STRP-ed definition of Structured Tandem Repeats in Proteins

Alexander Miguel Monzon[a,1], Paula Nazarena Arrías[b,1], Arne Elofsson[c], Pablo Mier[d], Miguel A. Andrade-Navarro[d], Martina Bevilacqua[b], Damiano Clementel[b], Alex Bateman[e], Layla Hirsh[f], Maria Silvina Fornasari[g], Gustavo Parisi[g], Damiano Piovesan[b], Andrey V. Kajava[h], Silvio C. E. Tosatto[b,*]

[a] Dept. of Information Engineering, University of Padova, via Giovanni Gradenigo 6/B, 35131 Padova, Italy
[b] Dept. of Biomedical Sciences, University of Padova, via U. Bassi 58/b, 35121 Padova, Italy
[c] Dept. of Biochemistry and Biophysics and Science for Life Laboratory, Stockholm University, Tomtebodavägen 23, 171 21 Solna, Sweden
[d] Institute of Organismic and Molecular Evolution, Faculty of Biology, Johannes Gutenberg University of Mainz, Hanns-Dieter-Hüsch-Weg 15, 55128 Mainz, Germany
[e] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[f] Dept. of Engineering, Faculty of Science and Engineering, Pontifical Catholic University of Peru, Av. Universitaria 1801 San Miguel, Lima 32, Lima, Peru
[g] Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, CONICET, Bernal, Buenos Aires, Argentina
[h] Centre de Recherche en Biologie cellulaire de Montpellier (CRBM), UMR 5237 CNRS, Université Montpellier, 1919 Route de Mende, Cedex 5, 34293 Montpellier, France,

ARTICLE INFO

Keywords:
Structured tandem repeats
Protein repeats
TRP dichotomies

ABSTRACT

Tandem Repeat Proteins (TRPs) are a class of proteins with repetitive amino acid sequences that have been studied extensively for over two decades. Different features at the level of sequence, structure, function and evolution have been attributed to them by various authors. And yet many of its salient features appear only when looking at specific subclasses of protein tandem repeats. Here, we attempt to rationalize the existing knowledge on Tandem Repeat Proteins (TRPs) by pointing out several dichotomies. The emerging picture is more nuanced than generally assumed and allows us to draw some boundaries of what is not a "proper" TRP. We conclude with an operational definition of a specific subset, which we have denominated STRPs (Structural Tandem Repeat Proteins), which separates a subclass of tandem repeats with distinctive features from several other less well-defined types of repeats. We believe that this definition will help researchers in the field to better characterize the biological meaning of this large yet largely understudied group of proteins.

1. Introduction

Since the advent of modern biochemistry, a considerable amount of work has been devoted to understanding proteins in all their different facets. Classifications have been established to describe proteins at the evolutionary (Mistry et al., 2021; Blum et al., 2021) and structural (Sillitoe et al., 2021) level. While these work well for globular proteins, the underlying domain paradigm has increasingly shown its limitations (Parisi et al., 2021). A particular area of difficulty are tandem repeat proteins (TRPs). Tandem repeats in proteins are generally defined as contiguous stretches of duplicated amino acid sequences (Kajava and Tosatto, 2018).

TRPs have been studied in terms of sequence - structure relationship (Kobe and Kajava, 2000). A strong dependence of the protein fold on the length of the sequence repeat has been used to build a classification of TRPs (Kajava, 2012). Similarly, studies trying to establish the evolutionary history of TRPs have found duplication events as the predominant mechanism for TRP expansion (Lang et al., 2000). At the same time, most of the functional knowledge about TRPs is limited to a handful of protein families, most notably solenoids (Kobe and Kajava, 2000; Andrade et al., 2001). It is currently unclear how far these concepts can be generalized. We argue that this is largely due to the presence of variability within multiple TRP attributes, which make it difficult to place TRPs all in the same box. In the following, we will try to address each of these as dichotomies, even though we are aware that not all of them can be simplified to either one of two options, before presenting an attempt to derive an operational definition of TRPs that encompasses their most salient features.

[*] Corresponding author.
E-mail address: silvio.tosatto@unipd.it (S.C.E. Tosatto).
[1] These authors have equally contributed to this work.

https://doi.org/10.1016/j.jsb.2023.108023
Received 29 April 2023; Received in revised form 31 July 2023; Accepted 28 August 2023
Available online 29 August 2023

Table 1
TRP sequence dichotomies. Examples of TRPs for each dichotomy.

Feature	Dichotomy	Sample protein	UniProtKB accession number
DNA	Repetitive	TBP_HUMAN; PolyQ in MEF2A_HUMAN	P20226; Q02078
	Non-repetitive	LPTA_ECOLI; E0RU15_SPITD	P0ADV1; E0RU15
Complexity	Low	S in PolyQ in HD_HUMAN PolyF193B_BOVIN	P42858; A7MB40
	High	GBB1_BOVIN; Q48391_KLEOX	P62871; Q48391
Conservation	High	LST8_HUMAN; PUM5_ARATH	Q9BVC4; Q9LJX4
	Low	TR10C_HUMAN; FIP2_ARATH	O14798; Q9SE95
Exons	Matching repeats	RINI_MOUSE; RCC1_HUMAN	Q91VI7; P18754
	Variable	COPG1_MOUSE; PEX5_HUMAN	Q9QZE5; P50542

2. TRP dichotomies

2.1. Sequence dichotomies

The first and most obvious dichotomy in TRPs is at the sequence level. Here, we can establish differences at the DNA level as well as repeat evolution and protein coverage. Table 1 summarizes TRP sequence dichotomies and provides examples of repeat proteins representing each feature.

2.1.1. Protein repeats at the DNA level

Repeats at the DNA level are very common. A recent study puts the fraction of the repetitive human genome at 53.9% of the genome sequence (Hoyt et al., 2022). Initially thought of as “junk”, repetitive DNA regions have been recognized as important sites for innovation in genomes (Makałowski, 2000), carrying out various important structural and functional roles (Shapiro and von Sternberg, 2005). Repetitive DNA can be further distinguished into tandem and interspersed repeats, with many sub-classes captured in the manually curated Dfam database (Storer et al., 2021).

Protein repeats are not necessarily translated from repeats at the DNA level. The degeneracy of the genetic code implies that even the most simple repeats, homorepeats (stretches of consecutive repetitions of a single amino acid (Jorda and Kajava, 2010), also known as polyX (Mier et al., 2017)), are not always codified by pure codon repeats. However, in most cases, the DNA replication slippage mechanism, causative of the generation of homorepeats, produces DNA repeats that simply translate to homorepeat sequences (Jorda and Kajava, 2010). Examples of this dichotomy are found in two adjacent homorepeats in human protein MEF2A (UniProtKB acc.:Q02078): polyQ (positions 420–430) is translated from DNA repeat [CAG]11, while polyP (positions 431–435) is translated from [CCG]2[CCA]2[CCG]1 (GenBank acc.: X68505.1). Synonymous point mutations within DNA repeats translating to homorepeats have been described to be evolutionarily selected in some cases to hamper the DNA slippage mechanism and to maintain a controlled non-toxic length (Rolfsmeier and Lahue, 2000; Mier and Andrade-Navarro, 2018). Longer and less perfect repeats forming 3D structures do not necessarily translate from repeated DNA. For example, TIM-barrels have evolved by gene duplication and fusion (Lang et al., 2000), and are assembled by repeats at the secondary structure level, in this case, and an eightfold repeat of (alpha/beta) units (Wierenga, 2001).

2.1.2. Sequence complexity

At the amino acid level, regions with tandem repeats can be formed by sequences of very different complexity. Sequence complexity is defined as the divergence of the residues forming a sequence or region from the expected background in a complete dataset in terms of usage and periodicity. There is a full range of examples from low to high complexity, from homorepeats to globular proteins.

The shorter the repeats the lower the complexity of the resulting sequence. Repeats of length one exemplify this perfectly. Homorepeats (polyX) represent the sequence with the lowest possible complexity. Repeats of short lengths bias the sequence as well because the repetition of the pattern only allows for a few amino acids. Only when the repeat length is large enough to accommodate a rich composition of amino acids can repeats eventually be formed by complex sequences.

The complexity of the repeats suggests disorder. It has been observed that perfect repeats tend to be disordered (Jorda et al., 2010). This is because perfect repeats reflect a recent duplication event, and generally, repeat duplications are easier to accommodate in disordered regions. However, when the lengths of the repeats become very short (and complexity reduces), this correlation vanishes because the structural properties of the repeat region depend more on the properties of the dominant amino acid. For repeats of length one, even though they all have the same level of sequence complexity, the polyQ region of human Huntingtin (UniProtKB acc.:P42858) has alpha-helical conformation (Urbanek et al., 2020), while the polyS region of the bovine protein FAM193B (UniProtKB acc.:A7MB40) is disordered.

Repeats with high complexity assemble to form entire domains, such as the beta-barrel structure in which protein CymA (UniProtKB:Q48391) folds into (PDB code: 4D51, van den Berg et al., 2015), or the beta-propeller formed by WD repeats in the bovine protein GNB1 (UniProtKB acc.:P62871), (PDB code: 1A0R; Loew et al., 1998). In these two examples, each of the repeats is long enough (about 40 amino acids), so they can have a rich amino acid composition and gain beta structure.

2.1.3. Conservation

Tandem repeat variation by means of unit gain/loss has been proposed as a source of the genetic variability needed for fast adaptation (Marcotte et al., 1999), which implies that conserved repeats might be a telltale sign of functional importance. Sequence conservation between units of the same protein has been proposed to infer events of unit gain through duplication of a single unit (Björklund et al., 2006).

In terms of repeat conservation across species, there seems to exist a variability in the degree of global conservation, which is organism-dependent. A study about human tandem repeat evolution by unit phylogeny (Schaper et al., 2014), shows that 61% of human repeats are conserved, i.e. there is no evidence of unit acquisition/loss, at least from the root of mammals, while perfect separation occurs in only a few cases and seems to be more associated with certain types of repeats, such as zinc-finger proteins.

However, in plants, although repeats are also highly conserved, there is evidence of highly mutable tandem repeats, and interestingly, these sometimes appear on proteins that are involved in pathogen defense, such as some LRRs belonging to R genes (Schaper and Anisimova, 2015).

2.1.4. Exon-bordered and not

In Eukaryotes, it has been suggested that the evolution of tandem repeat folds could be driven by exon duplication and reshuffling (Street et al., 2006; Björklund et al., 2006; Schaper and Anisimova, 2015). In this scenario, a single exon encodes for one structural repeat, thus, in cases of duplication or shuffling the new segment is usually structurally compatible with the old one. Proteins where multiple consecutive exons of similar length encode for structures that are similar and symmetric (Haigis et al., 2002; Light et al., 2012) support the hypothesis of evolution through exon duplication (Paladin et al., 2020), but no unique conclusion can be drawn for all repeat folds (Street et al., 2006; Schaper et al., 2014) as this is the case for some repeat families and not for others.

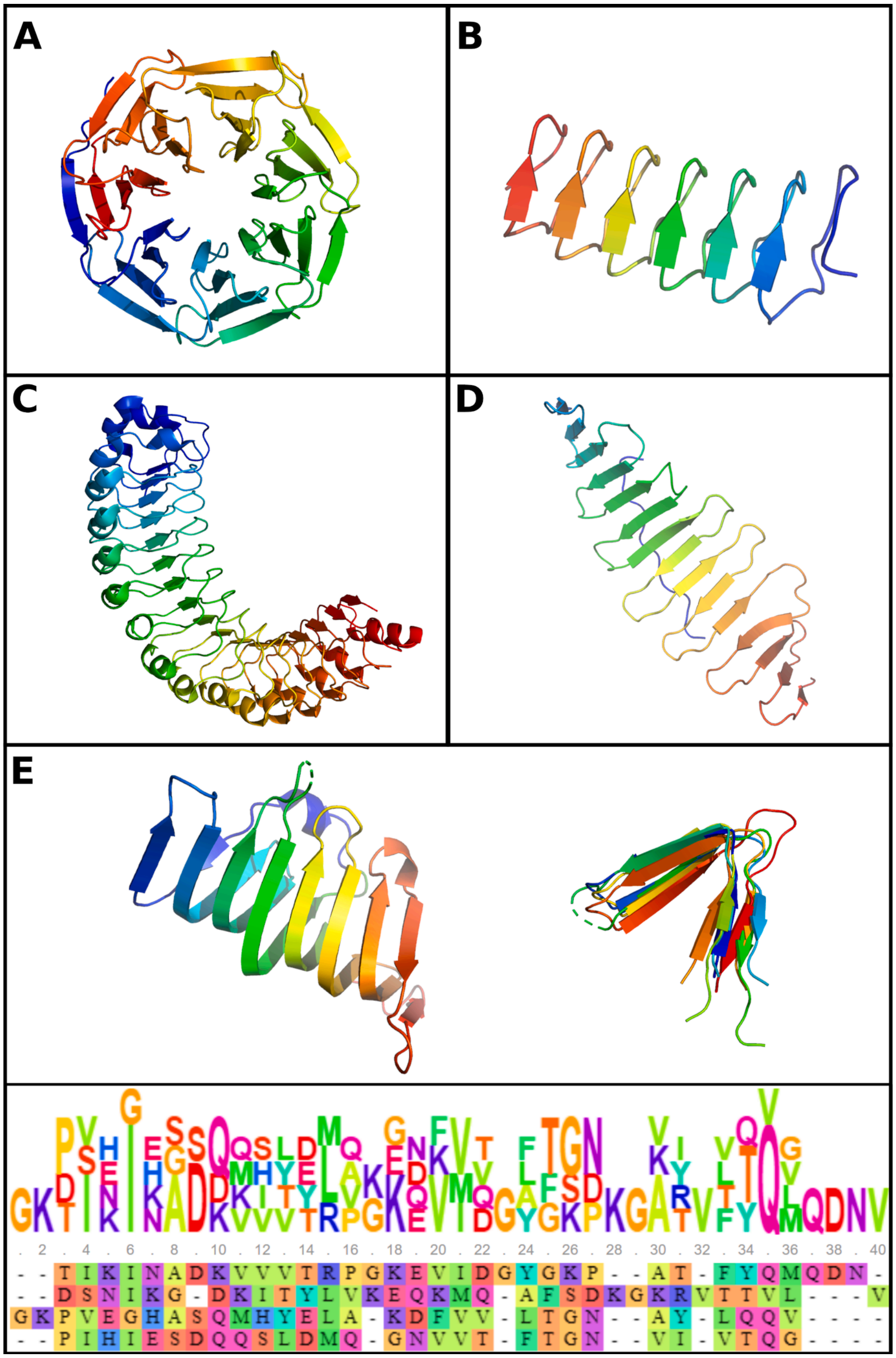
A
B
C
D
E
. 2 . 4 . 6 . 8 . 10 . 12 . 14 . 16 . 18 . 20 . 22 . 24 . 26 . 28 . 30 . 32 . 34 . 36 . 38 . 40
- - T I K I N A D K V V V T R P G K E V I D G Y G K P - - A T - F Y Q M Q D N -
- - D S N I K G - D K I T Y L V K E Q K M Q - A F S D K G K R V T T V L - - - V
G K P V E G H A S Q M H Y E L A - K D F V V - L T G N - - A Y - L Q Q V - - - -
- - P I H I E S D Q Q S L D M Q - G N V V T - F T G N - - V I - V T Q G - - - -

Fig. 1. Examples of shapes adopted by domains of tandem repeats. Identifiers, definition and classification correspond to entries in RepeatsDB. (A) Closed ring. 5eamA. WD repeat-containing protein 5 (aa 47–333), 7 units. Hierarchical classification: 4 Closed repeats, 4 Propeller, 1 Beta propeller with flat blades, 1 WD40 repeat. (B) Straight rod. 1ezgA. Crystal structure of antifreeze protein from the beetle, Tenebrio molitor (aa 8–79), 6 units. Hierarchical classification: 3 Elongated repeat, 1 Beta-solenoid. (C) Curved rod. 4u09A. Lic12759 (aa 44–417), 16 units. Hierarchical classification: 3 Elongated repeat, 2 Alpha/beta solenoid, 1 High curvature alpha/beta solenoid, 1 Leucine Rich Repeat. (D) Twisted rod. 4by2A. Spindle assembly abnormal 4 (Sas-4) (aa 73–203), 7 units. Hierarchical classification: 3 Elongated repeat, 4 Beta hairpins, 2 Single-layer beta-hairpin, 5 High curvature single-layer beta-hairpin. (E) Structure with imperfect structural alignment. 2r19B. Protein yhbn (aa 35–165), 4 units. Hierarchical classification: 3 Elongated repeat, 4 Beta hairpins, 2 Single-layer beta hairpin, 1 Beta-burrito. The level of sequence identity between the repeat units in E is similar to or higher than the examples shown in panels A-D, but the structures of the repeat units are more variable (e.g., the length of the beta strands changes from unit to unit).

In line with this observation, it was established that structural domains tend to be exon-bordered while disordered regions do not (Smithers et al., 2019). In terms of stability, repeat domains fall in-between these two categories due to their cooperative folding pathways. Even within the same repeat topology (Paladin et al., 2021), α-solenoids, some repeat families, such as Ankyrin repeats, were demonstrated to be exon-bordered while others such as Tetratricopeptide repeats, are encoded by a unique exon spanning the entire repeat array (Paladin et al., 2020). Other cases of repeat proteins show a complex repeat/exon pattern where the phasing is maintained, but different exons correspond to a different number of repeats (Björklund et al., 2010). Comparative analysis of the known isoforms generated by alternative splicing of TR-containing proteins shows that, as a rule, their different regions have an integer number of repeats and these changes do not cause serious perturbations in the overall 3D structure (Osmanli et al., 2022). At the same time, these structural changes create patches of new surfaces that can lead to the modification of protein functions.

2.2. Structure dichotomies

2.2.1. Flexibility

There have been reports of an overlap between proteins predicted to contain intrinsically disordered regions and proteins with tandem repeats. One reason is that tandem repeats are enriched in disorder-promoting residues, such as Gly and Ser (Delucchi et al., 2020). This is especially true for homorepeats, predominantly composed of hydrophilic and small residues, while large aliphatic, aromatic residues and Cys are rare (Chavali et al., 2020; Jorda and Kajava, 2010). Analysis of the human proteome in Swiss-Prot reveals that structured and unstructured repeat proteins are different in terms of repeat unit length and the number of repetitions. Unstructured repeat proteins, detected by overlapping the repeat annotation by Swiss-Prot and the disorder prediction from MobiDB-lite (Necci et al., 2020), mostly feature units of very short length and are repeated several times. The length of disordered repeat units varies considerably, ranging from less than ten amino acids to more than 60. STRiPs instead, identified in the human Swiss-Prot proteome from the repeat structures database RepeatsDB (Paladin et al., 2021), usually have longer (between 20 and 60 residues long) and the number of repetitions corresponding to the different length does not vary as much as for disordered repeats. These observations are coherent with the structural classification of repeats proposed in (Kajava, 2012). Furthermore, it was shown that increasing perfection of tandem repeat repetitiveness correlates with a stronger tendency to be unstructured (Jorda et al., 2010). There are interesting counter-examples such as the microbial cell surface protein SasG, which has perfect repeats of G5 and E domains and, from a sequence perspective, are predicted as disordered repeats. Although they are unstable in isolation, these repeats fold cooperatively into globular domains when expressed in tandem arrays (Gruszka et al., 2015).

One highly disordered repeat protein with low complexity is Mucin-16, which forms a protective barrier in the mucosa. It is also used as the antigen in a serum assay for monitoring ovarian epithelial cancer. The protein consists of one long extracellular region rich in Thr and Ser, which is heavily glycosylated, a single transmembrane domain and an intracellular domain. The extracellular region contains 16 SEA (Pfam ID: PF01390) domains linked by low complexity/disordered regions.

Another protein with repeated low complexity regions is the Huntingtin (HD_HUMAN) protein (Martin et al., 2014). This protein has one 21 residue long polyQ region at the N-termini, and two shorter polyP regions. In addition, it contains more than 30 HEAT tandem repeats. The structure of this protein has been solved in complex with HAP1 (Guo et al., 2018).

A repeat protein where the lines between structure and disorder are blurry is nebulin (Björklund et al., 2010; Yuen and Ottenheijm, 2020). Nebulin has a very interesting repeat structure, where seven domains, often covering four exons, are repeated in tandem (Björklund et al., 2010). These seven domains comprise one binding domain to an actin filament. The human nebulin protein contains about 35 of these seven-domain repeats, while other proteins, such as nebulette, contain fewer repeats. The nebulin protein is not predicted to be disordered, and secondary structure predictions predict one or two helices per domain. However, no nebulin structure is available and attempts to predict the structure have proven fruitless (Vlassi et al., 2013).

A number of proteins with tandem repeats fold in stable 3D structures (Wierenga, 2001; Björklund et al., 2006; Paladin et al., 2021). Typically, these proteins have long (more than ten residues) and imperfect repeats with high-complexity sequences.

2.2.2. *Shape*

Most proteins with aperiodic sequences have globular shapes. In contrast, proteins containing tandem repeats fold into structures having a variety of different shapes, which includes, in addition to globular shapes, elongated shapes and ring-like shapes (Fig. 1A). The elongated molecules, which are predominantly formed by solenoid structures, are especially varied in shapes ranging from straight (Fig. 1B) and twisted (Fig. 1C) to curved (Fig. 1D) rods (Paladin et al., 2021). Among these varieties of shapes, one can distinguish two major arrangements: (1) an open arrangement with, in principle, an unlimited number of repeats and (2) a closed one with a defined number of repeats. A typical example of an open shape would be a beta-solenoid and of closed one would be a beta-propeller (Fig. 1A).

Table 2
TRP structure dichotomies. Examples of TRPs for each dichotomy.

Feature	Dichotomy	Sample protein	UniProtKB accession number
Flexibility	Ordered	WDR61_HUMAN; SPTA1_HUMAN	Q9GZS3; P02549
	In between (secondary structure)	Nebulin (TAU_HUMAN); SF3B1_HUMAN	P10636; O75533
	Disordered	NUP2_YEAST; LEA1_APHAV	P32499; Q95V77
Shape	Open	TLR8_HUMAN; IMA1_MOUSE	Q9NR97; P52293
	Closed	RAG2_MOUSE; MANI_MARM1	P21784; F2JVT6

Length	Short	polyQ in HD_HUMAN;
		polyH in FA76B_HUMAN;
			P42858;
			Q5HYJ3;

	Long	DAF_HUMAN; NECT3_HUMAN	P08174; Q9NQS3
Folding	Non-independent	PUF4_YEAST; CDN2A_HUMAN	P25339; P42771
	Independent	MCP_HUMAN; FINC_HUMAN	P15529; P02751

2.2.3. *Long and short repeats*

Analysis of the known 3D structures of proteins with repeats suggested a classification that subdivides them based on their repeat length and this classification allows streamlining this high diversity of the structures (Kajava, 2012; Paladin et al., 2021). Based on repeat length, the structures with repeats can be broadly divided into five classes. Class I includes repeat lengths of 1–2 residues, and these proteins can form crystalline aggregates. Class II with repeat units of 3–5 residues covers fibrous structures such as collagen or alpha-helical coiled-coils; Class III – elongated structures where the repetitive units require each other for structural stability such as solenoid proteins; Class IV – closed repetitive structures, which include TIM-barrels and b-propellers and Class V – bead on a string structure.

2.2.4. *Folding*

Short tandem repeats with 1 or 2 residue units (Class I) are frequently composed of hydrophilic amino acids coding for intrinsically disordered regions. However, when these tandem repeats have potential to form 3D structures, they self-assemble together in aggregates (Jorda and Kajava, 2010). Known fibrous proteins (Class II), such as collagen or α-helical coiled-coils, fold into regular structures upon oligomerization. In some cases, for example for collagen, folding represents a multistep process. This behaviour is functionally relevant, allowing the collagen triple-helix to be formed only after transportation through the cytoplasm to the extracellular matrix (Ishikawa and Bächinger, 2013).

The folding paradigm for Class III TRPs, especially solenoids, has been addressed by different authors in the literature (Kajander et al., 2005; Barrick et al., 2008). Different studies, both computational and experimental, show that the folding mechanisms of these proteins are very diverse and cannot be generalized (Galpern et al., 2022; Espada et al., 2015). For example, both the Ankyrin domain of *Drosophila melanogaster* Notch protein and the LRR of Internalin B of *Lysteria monocytogenes* display fully cooperative (all-or-none transitions) folding behaviour (Courtemanche and Barrick, 2008; Bradley and Barrick, 2002). However, within the LRR domain of PP32, the folding process begins with the C-terminus, potentially acting as a catalyst for the subsequent folding of the remaining structure (Dao et al., 2014). There is also evidence of parallel folding pathways for consensus ankyrin repeats in which the nuclei of folding can start in different places, increasing the folding rate with chain length (Aksel and Barrick, 2014). Closed (Class IV) TRPs are also thought to fold in a similar way. Beads on a string TRPs (class V) are different, in that they are composed of consecutive globular domains. From a folding perspective, these TRPs can be both structurally

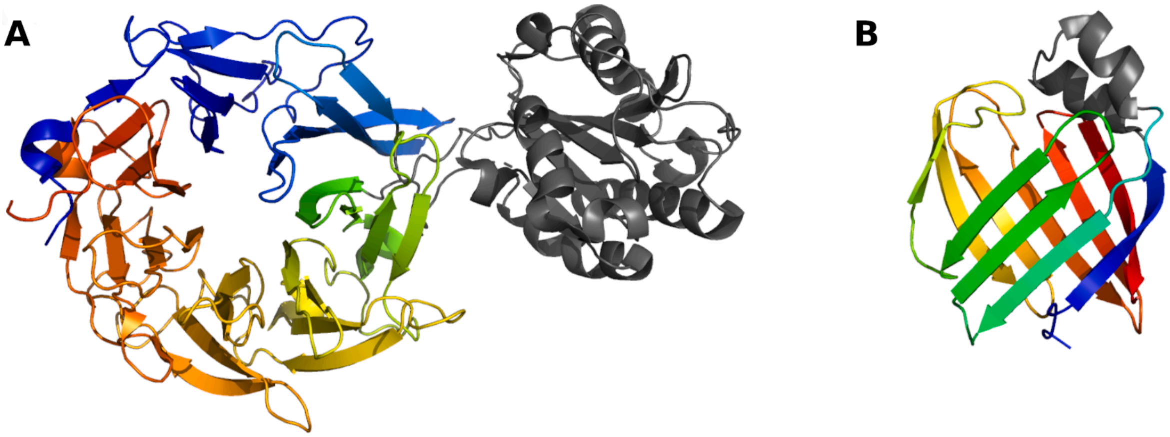

Fig. 2. Insertions. A) Propeller of Integrin alpha-L (P20701, PDB code: 5E6U) with a von Willebrand factor type A domain (in dark gray) inserted between its blades. B) Beta-barrel of Retinol-binding protein 2 (P50120, PDB code: 4ZGU) with an alpha-hairpin insertion (in dark gray).

autonomous, i.e. they can fold individually and do not require the neighboring repeats for stabilization, or they might be able to fold individually but are stabilized through inter-domain interactions, as for example in the Spectrin protein (Petersen and Barrick, 2021; Batey and Clarke, 2006).

All structure-related dichotomies are summarized in Table 2.

2.3. Repetitiveness dichotomies

2.3.1. Protein coverage

Protein structural tandem repeats fold into repeated structural units that pack against each other forming compact domains (Andrade et al., 2001). As such, one could say that the properties that apply in general to domains being units of evolution, function and structure apply also to domains formed by tandem repeats. Therefore, it is not too difficult to find examples of proteins entirely formed by a domain of tandem repeats, multiple domains with only some being tandem repeats, and even those combining more than one domain formed by tandem repeats (either of the same type or of different types). A particularity of these domains is their flexibility and evolutionary adaptability, which made them very appropriate for interactions with other proteins. The ribonuclease inhibitor is a protein of around 450 amino acids entirely composed of Leucine-rich repeats (LRRs) and its structure solved in complex with ribonuclease A shows how the beta-sheet formed by consecutive repeats coils around the targeted protein (Kobe and Deisenhofer, 1995). Huntingtin is an example of a very large multidomain protein that alternates several domains composed of HEAT repeats, covering approximately 75% of the sequence, with non-repeat-containing domains (Guo et al., 2018). In such a protein, the function of flexible tandem repeat domains, protein interaction, is dominant. Huntingtin is known to interact with hundreds of proteins, in a multiplicity of interaction sites modulated by post-translational modifications and compositionally biased regions (Harjes and Wanker, 2003; Kastano et al., 2021). In contrast, there are large proteins with a small fraction composed of tandem repeats, where protein interaction function is not dominant. One example is the mineralocorticoid receptor, which contains a domain of about 150 amino acids composed of repeats of approximately 10 amino acids (Vlassi et al., 2013). This protein has other domains, for example, to interact with DNA for transcriptional control and to respond to a hormone, and the structure of these globular domains is known but not that of the tandem repeats. In this case, the tandem repeat domain holds many serine residues that are the target of phosphorylation. It was hypothesized that the domain is structured in the de-phosphorylated state and becomes unstructured upon phosphorylation. In this case, the tandem repeat domain could have a function in protein interaction regulated by phosphorylation.

Evolutionary analyses suggest that, while tandem repeats ensembles are propitious to increase or decrease by addition of extra repeats or by their removal, respectively, examination of homologs suggests little flexibility once the tandem repeat domain is established (Kamel et al., 2021). Therefore, the examples above of proteins that include tandem repeat domains are stable when different family members are compared, with respect to the fraction of the protein covered by tandem repeats.

2.3.2. Variability in the number of repeats

There are repeats that form elongated structures (e.g. TPRs, LRRs, alpha-solenoids, etc). Those have great variability in the number of repeat units since there is no constraint of the length they can reach. In contrast, repeats that form closed structures (like barrels or propellers) have a tendency to be much more constrained in the number of units. This feature is so definitory that in theory, even in the absence of any structural information, it should be possible to know if a repeat is forming open or closed structures by analyzing the distribution of the number of repeat units observed in many proteins: repeats forming open structures have a distribution in the number of repeats that decreases continuously versus the number of units, whereas repeats forming closed structures tend to occur in multiples of the canonical number of units of the repeat domain since sometimes more than one of these closed structures can occur within a protein sequence (Andrade et al., 2001).

2.3.3. Insertions

The modularity and symmetry of tandem repeat structures can be sometimes interrupted by non-repetitive segments. These segments are commonly referred to as “insertions”, and can be found either inside or in-between units. Structurally, insertions serve no purpose in maintaining the stability of the repeat fold, but instead, they might provide the ability to perform specific functions, such as binding. For example, some alpha integrins have a von Willebrand factor type A domain inserted in their propeller regions, which contains a metal-ion-dependent adhesion site (MIDAS) (Takada et al., 2007). Another example of functional insertions inside a repeat unit is the insertion of a helix or a bigger domain in beta-barrels, which regulate their opening and closing, like in the case of Retinol-binding protein 2 (P50120, PDB code: 4ZGU) (Chen et al., 1998) or the bacterial porin oprP (P05695, PDB code: 2O4V) (Fig. 2).

Smaller insertions, like those composed of longer loops, can also be functional. For example, it has been proposed that the longer loops present in the beta-helix protein Pertactin of *Bordetella pertussis* can help in immune evasion by epitope masking (Hijnen et al., 2007).

2.3.4. Repeat perfection

Proteins that are able to fold and have the same amino acid sequence

Table 3
TRP repetitiveness dichotomies. Examples of TRPs for each dichotomy.

Feature	Dichotomy	Sample protein	UniProtKB accession number
Coverage	Complete	RINI_PIG; A0A0M3KL00_STRMG	P10775; A0A0M3KL00
	Partial	TLR3_HUMAN; RAG2_MOUSE GLMU_MYCTU	O15455 (PDB code:7C76); P21784 P9WMN3
Number	Fixed number	OMPG_ECOLI; TPIS_PLAFA	P76045; Q07412

	Variable number	IPO13_HUMAN;
		B9MKT4_CALBD;
			O94829;
			B9MKT4;

Insertions	Insertion-less	E7FCY1_DANRE; BBKI_BAUBA	E7FCY1; P83052
	Insertion-prone	Q3JR17_BURP1; NANA_STRR6	Q3JR17; P62576
Perfection	High	Q3ZD72_XANCA; TAL2_TACTR	Q3ZD72; Q27084
	Low	WDR5_HUMAN; CEEP_RHOMR	P61964; F8WRK9

usually adopt the same 3D structure. Similarly, if a protein has an array of perfect tandem repeats in its sequence, i.e. every repeat has exactly the same sequence, each repetitive unit usually has the same 3D structure (examples, PDB code: 1NA0 designed TPR structure and PDB code: 3UGM TAL effector with HEAT-like repeats and PDB code: 4YCW, *P. anserina* TPR containing protein (Marold et al., 2015)). Repeat perfection can be estimated by a parameter called P_{sim} (Jorda et al., 2010), in which $P_{\text{sim}} = 1$ represents total perfection, and values of $0.7 \leq P_{\text{sim}} < 1$ represent nearly perfect repeats. However, protein tandem repeats are frequently not perfect, containing a number of mutations (substitutions, insertions, deletions) accumulated during evolution. In reality, naturally occuring perfect tandem repeats seem to be very rare, with only a very low percentage of them having a $P_{\text{sim}} = 1$ (Jorda et al., 2010), and most of which adopt disordered conformations. Most of the time, the decrease of the amino acid sequence similarity between repeat units leads to the increase of the difference in their 3D structure, but the conservation of specific amino acids highlights their importance for the maintenance of the repeat fold. (Example, PDB code: 1DAB, pertactin). However, an interesting case is the Pumilio family (PFAM Identifier: PF0806), also known as PUF repeats, composed of proteins that regulate translation and mRNA stability of eukaryotic organisms including mammals, flies, worms, slime mold, and yeast (Zamore et al., 1997). This is a good example of structure perfection and sequence degeneration. On one hand, the repeat units show a high structural similarity (average TM-Score of 0.75) and, on another hand, the average sequence identity between units is 26 percent (Supplementary Fig. 1).

Repetitiveness-related dichotomies are summarized in Table 3.

3. Discussion

3.1. Repeat evolution

By tracing the similarity between repeat units it is possible to learn more about the evolutionary processes for a particular repeat family. At the moment of duplication, two repeats are identical both at DNA and amino acid levels, but then by time, they will diverge. This means that two repeats that are more similar have a more recent ancestry than repeats that are more different. When comparing the similarity of repeats, it was found that for most families adjacent repeats are, on average, less similar to each other than repeats separated by a few repeats (Björklund et al., 2006). However, repeats further away are even less similar, and the exact patterns vary from family to family. The most intriguing family is nebulin, which almost exclusively seems to be duplicated in groups of seven repeats (Björklund et al., 2010). We have interpreted this as the most likely mechanism is a tandem duplication that favours duplication of identical (or highly similar) DNA segments.

3.2. Structure/unstructured

One of the first questions to answer when approaching a repetitive protein to predict its 3D structure is whether it is structured or not. Today, thanks to a number of methods for prediction of IDRs we can answer this question with some accuracy (Necci et al., 2021). In addition, proteins with tandem repeats have several supplementary correlations, which can improve this prediction. For example, the increasing perfection of repetitiveness correlates with a stronger tendency of tandem repeats to be unstructured (Jorda et al., 2010). Our analysis based on Human repeat proteins in Swiss-Prot also suggests that the shorter the length of the repeats, the less structured the tandem repeats are. At the same time, the shorter the repeat length, the lower is their complexity (Mier et al., 2020). Thus, we can also assert that the tandem repeats with low-complexity sequences are most probably unstructured.

3.3. Future perspectives repeats and structure prediction

In December 2020 DeepMind demonstrated at CASP14 that the structure of virtually all, well folded single domain, proteins can be predicted at an accuracy close to what can be obtained experimentally (Laine et al., 2021; AlQuraishi, 2021). For a long time, it was believed that repeat proteins caused special challenges for co-evolution based structure prediction methods. However, recently it was shown that the latest, deep-learning-based methods, can predict the structure of many repeat proteins (Bassot and Elofsson, 2021). In this study, we show that about 90% of the repeat proteins can be accurately predicted. Further, we also provide novel, most likely correct, multiple repeat models for 41 out of 48 PFAM families lacking a protein structure. The seven families where no reliable structure is predicted include two families with short motifs C_tripleX (PF02363) and Lipoprotein_15 (PF03640), (Baker model), two membrane-associated families SVM_repeat (PF13753), Ish1 (PF10281), two short families, Nebulin (PF00880), INT_rpt (PF14882) and one longer the WD repeat and coiled-coil-containing protein family (PF15390). Models for all these seven families are available from Pfam for single repeat units.

As mentioned above many repeat proteins contain disorder, if these proteins have a structure in some physiologically relevant environments and if such structures can be predicted remains to be seen.

Likely the latest progress by DeepMind will make these early attempts to model these repeat proteins outdated, once they are applied to repeat proteins en masse. It is likely that the main challenge will not be to model the individual proteins but to model the interaction of repeat proteins. Here some progress has been made both for homomeric (Quadir et al., 2020) and heteromeric interactions (Baek et al., 2021). However, these have not yet been applied to repeat proteins, but we see no major reason for these strategies not to work on at least some repeat proteins, promising a bright future for the structural analysis of repeat proteins.

The regularity, modularity and linearly arrayed structures of TRPs make them attractive targets for protein design (Javadi and Itzhaki, 2013; Brunette et al., 2015). However, there are several challenges in protein design involving TRPs such us inefficient self-assembly, control of repeat-protein curvature, stability, diverse inter-repeat geometry, folding kinetics and undesired higher-order structures that have to be addressed (Doyle et al., 2015; Hallinan et al., 2021; Brunette et al., 2015). The improvements in structure prediction methods have the potential to greatly facilitate the design of tandem repeat proteins with desired properties. This could lead to more efficient design processes and the development of novel proteins for various applications in biotechnology, medicine, and research. In recent work, proteins with repeating units were designed to bind peptides with repeating sequences, and geometric hashing was used to identify compatible protein backbones and peptide-docking arrangements (Wu et al., 2023).

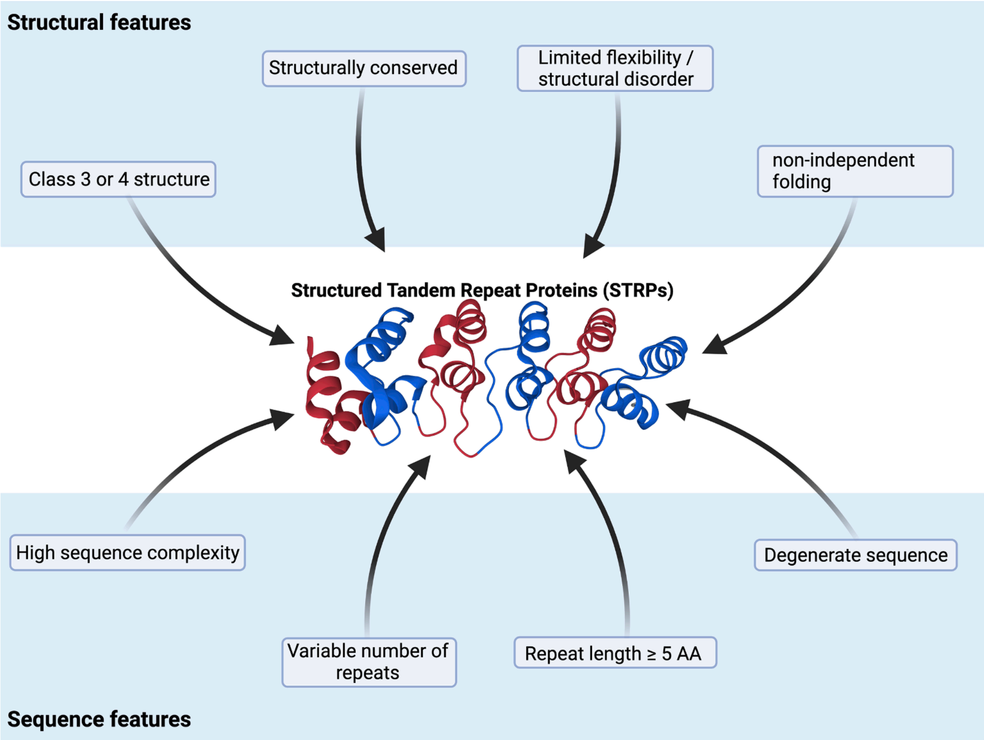

Fig. 3. Overview of STRP features. Single protein chain definition.

Furthermore, the progress made in natural language processing has enabled the development of protein language models, with ProGPT2 being one example (Ferruz et al., 2022). These models have the capability to produce sequences that exhibit distant relationships to natural protein sequences and possess structures that resemble those found within the known structural space, even including non-idealized complex structures.

3.4. *Improve repeat annotation*

As it has been mentioned throughout this work, repeat proteins present multiple dichotomies. These dichotomies represent a challenge when it comes to repeat annotation. RepeatsDB (Paladin et al., 2021) is a structured tandem repeat proteins database that serves as an important source of both manually and automatically annotated experimental protein structures for the scientific community. The database provides the user with information about start and end of regions, units and insertions, as well as classification. Currently, computational methods to automatically detect STRPs lack accuracy, and their long execution times defy their application on big structural databases. Consequently, manual curation of repeat protein structures is a time consuming task, which requires human inspection of every structure, even those belonging to the same protein. Expanding the annotation of the repeat folds is of utmost importance for developing and improving automatic detection methods, as well as for broadening the classification of repeats, among other tasks.

3.5. *Disease*

Since tandem repeats offer easy to evolve, large and flexible structures very appropriate for interacting with other proteins (Andrade et al., 2001), and many disease-causing mutations involve modifications of the protein interaction network (Yeger-Lotem and Sharan, 2015), it can be expected that many tandem repeat proteins will be associated with genetic diseases. However, there does not seem to be a significant enrichment of disease association in tandem repeat proteins. For example, a query of the human reviewed proteins in the current version of UniProt (2023_03) returns 20,423 proteins of which 5,006 are annotated as involved in disease (24.6%). For comparison, of the 1,957 human reviewed proteins annotated as having a repeat, 580 were involved in disease (29.6%), which is a higher value but not terribly different. More specifically, the fraction of disease-related proteins can be even lower for some types of tandem repeats, e.g. 62 of 312 (19.9%) proteins annotated as having LRR tandem repeats. One could

Table 4
Summary of TRP dichotomies. STRPs subset preferred features are in bold and underlined.

SEQUENCE		STRUCTURE		REPETITIVENESS	
Feature	Dichotomy	Feature	Dichotomy	Feature	Dichotomy
DNA	Repetitive	***Flexibility***	**Ordered**	*Coverage*	Partial
	Non-Repetitive		In between		Complete
Complexity	**High**		Disordered	***Number***	Fixed
	Low	*Shape*	Open		**Variable**
Conservation	High		Closed	*Insertions*	Insertion-less
	Low	*Length*	Short		Insertion prone
Exons	Matching		Long	***Perfection***	High
	Variable	***Folding***	**Non-independent**		**Low**
			Independent		

hypothesize that while the folding of some TRPs seems more unstable than that of globular proteins, this flexibility could allow them to accommodate more mutations, which is actually observed in the inter-repeat variability of many TRPs, which would mean them being less associated with genetic disease than other proteins. For example, 17 disease-causing mutations have been discussed recently in the ANKRD11 protein, and none of them fall in the domain of Ankyrin repeats contained by this protein (Parenti et al., 2021).

4. Conclusion

In this review we have shown how TRPs cover a wide range of different properties, making them difficult to pigeonhole into a single class with specific characteristics. Rather, TRPs should be seen as a mixed bag of different sub-phenomena which are partially contradicting each other. This plasticity is however in contrast with a subset of TRPs which can be considered “well-behaved”, in the sense that they share a common set of identifying features. This subset however is of interest, as it will allow us to draw some biological conclusions on the underlying phenomena. We will therefore now endeavour to define this subset, which we will tentatively call “structured tandem repeat proteins” or “STRPs”, as an operational definition (Fig. 3 and Table 4).

STRPs are tandem repeat proteins for which their structure can be solved with structural biology experiments, such as X-ray crystallography or electron microscopy, or likely well predicted by state-of-the-art structure prediction methods like AlphaFold (Jumper et al., 2021). STRPs have clear secondary structure propensities and form regular tertiary structures, which can be part of large molecular assemblies. Typically, these proteins fall into Classes III (i.e. elongated) and IV (i.e. closed) of Kajava’s classification of protein tandem repeats. They fold non-independently, but their folding patterns cannot be generalized beyond this characteristic. As such, there is little room for flexibility or intrinsic disorder, which is limited to partial folding of certain units as in the case of Ankyrin repeats (e.g. IKBalpha). The sequence features of STRPs are repeat units of at least five residues which also exhibit high sequence complexity. STRPs can be highly degenerate in sequence while maintaining a similar structure and exhibit a variable number of repeat units, suggesting a decoupling between structural size and protein function.

STRPs can be thought of as the complement to low complexity regions (LCRs) in proteins (Mier et al., 2020). Where the latter is defined primarily through a set of sequence features, STRPs are predominantly defined from their structural features. What both LCRs and STRPs have in common, is their difference in behaviour from regular globular domains. The operational definition proposed for STRPs opens the door for further research in order to better understand their function and biological implications.

CRediT authorship contribution statement

Alexander Miguel Monzon: Conceptualization, Formal analysis, Writing – original draft, Methodology, Data curation. **Paula Nazarena Arrías:** Conceptualization, Formal analysis, Writing – original draft, Methodology, Data curation. **Arne Elofsson:** Formal analysis, Writing – review & editing, Methodology. **Pablo Mier:** Formal analysis, Writing – review & editing, Methodology. **Miguel A. Andrade-Navarro:** Formal analysis, Writing – review & editing, Methodology. **Martina Bevilacqua:** Formal analysis, Methodology. **Damiano Clementel:** Formal analysis, Methodology. **Alex Bateman:** Conceptualization, Writing – original draft. **Layla Hirsh:** Formal analysis, Writing – review & editing, Methodology. **Maria Silvina Fornasari:** Conceptualization, Writing – original draft. **Gustavo Parisi:** Conceptualization, Writing – original draft. **Damiano Piovesan:** Conceptualization, Writing – original draft. **Andrey V. Kajava:** Supervision, Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing. **Silvio C.E. Tosatto:** Supervision, Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors are grateful to Lisanna Paladin for initial help in the manuscript, and Jiachen Lu for helping in the manual curation of RepeatsDB. This work was supported by European Union’s Horizon 2020 research and innovation programme under grant agreement No 823886 (H2020 MSCA-RISE “REFRACT”) and based upon work from COST Action ML4NGP, CA21160, supported by COST (European Cooperation in Science and Technology).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jsb.2023.108023.

References

Aksel, T., Barrick, D., 2014. Direct Observation of Parallel Folding Pathways Revealed Using a Symmetric Repeat Protein System. Biophys. J. 107, 220–232.
AlQuraishi, M., 2021. Machine learning in protein structure prediction. Curr. Opin. Chem. Biol. 65, 1–8.
Andrade, M.A., et al., 2001. Protein Repeats: Structures, Functions, and Evolution. J. Struct. Biol. 134, 117–131.
Baek, M., et al., 2021. Accurate prediction of protein structures and interactions using a 3-track network. bioRxiv, 2021.06.14.448402.
Barrick, D., et al., 2008. Folding landscapes of ankyrin repeat proteins: experiments meet theory. Curr. Opin. Struct. Biol. 18, 27–34.
Bassot, C., Elofsson, A., 2021. Accurate contact-based modelling of repeat proteins predicts the structure of new repeats protein families. PLOS Comput. Biol. 17, 1–20.
Batey, S., Clarke, J., 2006. Apparent cooperativity in the folding of multidomain proteins depends on the relative rates of folding of the constituent domains. Proc. Natl. Acad. Sci. 103, 18113–18118.
Björklund, A.K., et al., 2006. Expansion of protein domain repeats. PLoS Comput. Biol. 2, e114.
Björklund, A.K., et al., 2010. Nebulin: a study of protein repeat evolution. J. Mol. Biol. 402, 38–51.
Blum, M., et al., 2021. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 49, D344–D354.
Bradley, C.M., Barrick, D., 2002. Limits of Cooperativity in a Structurally Modular Protein: Response of the Notch Ankyrin Domain to Analogous Alanine Substitutions in Each Repeat. J. Mol. Biol. 324, 373–386.
Brunette, T.J., et al., 2015. Exploring the repeat protein universe through computational protein design. Nature 528, 580–584.
Chavali, S., et al., 2020. Amino acid homorepeats in proteins. Nat. Rev. Chem. 4, 420–434.
Chen, X., et al., 1998. Crystal structure of apo-cellular retinoic acid-binding protein type II (R111M) suggests a mechanism of ligand entry11Edited by I. A. Wilson. J. Mol. Biol. 278, 641–653.
Courtemanche, N., Barrick, D., 2008. Folding thermodynamics and kinetics of the leucine-rich repeat domain of the virulence factor Internalin B. Protein Sci. 17, 43–53.
Dao, T.P., et al., 2014. Capping motifs stabilize the leucine-rich repeat protein PP32 and rigidify adjacent repeats. Protein Sci. 23, 801–811.
Delucchi, M., et al., 2020. A New Census of Protein Tandem Repeats and Their Relationship with Intrinsic Disorder. Genes 11.
Doyle, L., et al., 2015. Rational design of α-helical tandem repeat proteins with closed architectures. Nature 528, 585–588.
Espada, R., et al., 2015. Repeat proteins challenge the concept of structural domains. Biochem. Soc. Trans. 43, 844–849.
Ferruz, N., et al., 2022. ProtGPT2 is a deep unsupervised language model for protein design. Nat. Commun., 13, 4348.
Galpern, E.A., et al., 2022. Evolution and folding of repeat proteins. Proc. Natl. Acad. Sci. 119, e2204131119.
Gruszka, D.T., et al., 2015. Cooperative folding of intrinsically disordered domains drives assembly of a strong elongated protein. Nat. Commun. 6, 7271.

Guo, Q., et al., 2018. The cryo-electron microscopy structure of huntingtin. Nature 555, 117–120.

Haigis, M.C., et al., 2002. Evolution of ribonuclease inhibitor by exon duplication. Mol. Biol. Evol. 19, 959–963.

Hallinan, J.P., et al., 2021. Design of functionalised circular tandem repeat proteins with longer repeat topologies and enhanced subunit contact surfaces. Commun. Biol. 4, 1–14.

Harjes, P., Wanker, E.E., 2003. The hunt for huntingtin function: interaction partners tell many different stories. Trends Biochem. Sci. 28, 425–433.

Hijnen, M., et al., 2007. The role of peptide loops of the Bordetella pertussis protein P.69 pertactin in antibody recognition. Vaccine 25, 5902–5914.

Hoyt, S.J., et al., 2022. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science 376, eabk3112.

Ishikawa, Y., Bächinger, H.P., 2013. A molecular ensemble in the rER for procollagen maturation. Biochim. Biophys. Acta 1833, 2479–2491.

Javadi, Y., Itzhaki, L.S., 2013. Tandem-repeat proteins: regularity plus modularity equals design-ability. Curr. Opin. Struct. Biol. 23, 622–631.

Jorda, J., et al., 2010. Protein tandem repeats - the more perfect, the less structured. FEBS J. 277, 2673–2682.

Jorda, J., Kajava, A.V., 2010. Protein Homorepeats: Sequences, Structures, Evolution, and Functions. In: Alexander McPherson (Ed.), Advances in Protein Chemistry and Structural Biology. Academic Press, pp. 59–88.

Jumper, J., et al., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589.

Kajander, T., et al., 2005. A new folding paradigm for repeat proteins. J. Am. Chem. Soc. 127, 10188–10190.

Kajava, A.V., 2012. Tandem repeats in proteins: from sequence to structure. J. Struct. Biol. 179, 279–288.

Kajava, A.V., Tosatto, S.C.E., 2018. Editorial for special issue "Proteins with tandem repeats: sequences, structures and functions". J. Struct. Biol. 201, 86–87.

Kamel, M., et al., 2021. REP2: A Web Server to Detect Common Tandem Repeats in Protein Sequences. J. Mol. Biol. 433, 166895.

Kastano, K., et al., 2021. The Role of Low Complexity Regions in Protein Interaction Modes: An Illustration in Huntingtin. Int. J. Mol. Sci. 22, 1727.

Kobe, B., Deisenhofer, J., 1995. A structural basis of the interactions between leucine-rich repeats and protein ligands. Nature 374, 183–186.

Kobe, B., Kajava, A.V., 2000. When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. Trends Biochem. Sci. 25, 509–515.

Laine, E., et al., 2021. Protein sequence-to-structure learning: Is this the end(-to-end revolution)?.

Lang, D., et al., 2000. Structural Evidence for Evolution of the β/α Barrel Scaffold by Gene Duplication and Fusion. Science 289, 1546–1550.

Light, S., et al., 2012. The evolution of filamin-a protein domain repeat perspective. J. Struct. Biol. 179, 289–298.

Loew, A., et al., 1998. Phosducin induces a structural change in transducin beta gamma. Structure 6, 1007–1019.

Makałowski, W., 2000. Genomic scrap yard: how genomes utilize all that junk. Gene 259, 61–67.

Marcotte, E.M., Pellegrini, M., Yeates, T.O., Eisenberg, D., 1999. A census of protein repeats. J. Mol. Biol. 293 (1), 151–160. https://doi.org/10.1006/jmbi.1999.3136. PMID: 10512723.

Marold, J.D., et al., 2015. A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins. Structure 23, 2055–2065.

Martin, D.D.O., et al., 2014. Identification of a post-translationally myristoylated autophagy-inducing domain released by caspase cleavage of huntingtin. Hum. Mol. Genet. 23, 3166–3179.

Mier, P., et al., 2017. Context characterization of amino acid homorepeats using evolution, position, and order: Characterization of Amino Acid Homorepeats. Proteins Struct. Funct. Bioinforma. 85, 709–719.

Mier, P., et al., 2020. Disentangling the complexity of low complexity proteins. Brief. Bioinform. 21, 458–472.

Mier, P., Andrade-Navarro, M.A., 2018. Glutamine Codon Usage and polyQ Evolution in Primates Depend on the Q Stretch Length. Genome Biol. Evol. 10, 816–825.

Mistry, J., et al., 2021. Pfam: The protein families database in 2021. Nucleic Acids Res. 49, D412–D419.

Necci, M., et al., 2020. MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavors in proteins. Bioinformatics.

Necci, M., et al., 2021. Critical assessment of protein intrinsic disorder prediction. Nat. Methods 18, 472–481.

Osmanli, Z., et al., 2022. The Difference in Structural States between Canonical Proteins and Their Isoforms Established by Proteome-Wide Bioinformatics Analysis. Biomolecules 12, 1610.

Paladin, L., et al., 2020. A novel approach to investigate the evolution of structured tandem repeat protein families by exon duplication. J. Struct. Biol. 212, 107608.

Paladin, L., et al., 2021. RepeatsDB in 2021: improved data and extended classification for protein tandem repeat structures. Nucleic Acids Res. 49, D452–D457.

Parenti, I., et al., 2021. ANKRD11 variants: KBG syndrome and beyond. Clin. Genet. 100, 187–200.

Parisi, G., et al., 2021. "Protein" no longer means what it used to. Curr. Res. Struct. Biol. 3, 146–152.

Petersen, M., Barrick, D., 2021. Analysis of Tandem Repeat Protein Folding Using Nearest-Neighbor Models. Annu. Rev. Biophys. 50, 245–265.

Quadir, F., et al., 2020. Predicting interchain contacts for homodimeric and homomultimeric protein complexes using multiple sequence alignments of monomers and deep learning. bioRxiv, 2020.11.09.373878.

Rolfsmeier, M.L., Lahue, R.S., 2000. Stabilizing effects of interruptions on trinucleotide repeat expansions in Saccharomyces cerevisiae. Mol. Cell. Biol. 20, 173–180.

Schaper, E., et al., 2014. Deep conservation of human protein tandem repeats within the eukaryotes. Mol. Biol. Evol. 31, 1132–1148.

Schaper, E., Anisimova, M., 2015. The evolution and function of protein tandem repeats in plants. New Phytol. 206, 397–410.

Shapiro, J.A., von Sternberg, R., 2005. Why repetitive DNA is essential to genome function. Biol. Rev. Camb. Philos. Soc. 80, 227–250.

Sillitoe, I., et al., 2021. CATH: increased structural coverage of functional space. Nucleic Acids Res. 49, D266–D273.

Smithers, B., et al., 2019. 'Why genes in pieces?'-revisited. Nucleic Acids Res. 47, 4970–4973.

Storer, J., et al., 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob. DNA 12, 2.

Street, T.O., et al., 2006. The Role of Introns in Repeat Protein Gene Formation. J. Mol. Biol. 360, 258–266.

Takada, Y., et al., 2007. The integrins. Genome Biol. 8, 215.

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

van den Berg, B., et al., 2015. Outer-membrane translocation of bulky small molecules by passive diffusion. Proc. Natl. Acad. Sci. U.S.A. 112, E2991–E2999.

Vlassi, M., et al., 2013. Short tandem repeats in the inhibitory domain of the mineralocorticoid receptor: prediction of a β-solenoid structure. BMC Struct. Biol. 13, 17.

Wierenga, R.K., 2001. The TIM-barrel fold: a versatile framework for efficient enzymes. FEBS Lett. 492, 193–198.

Wu, K., et al., 2023. De novo design of modular peptide-binding proteins by superhelical matching. Nature 616, 581–589.

Yeger-Lotem, E., Sharan, R., 2015. Human protein interaction networks across tissues and diseases. Front. Genet. 6, 257.

Yuen, M., Ottenheijm, C.A.C., 2020. Nebulin: big protein with big responsibilities. J. Muscle Res. Cell Motil. 41, 103–124.

Zamore, P.D., et al., 1997. The Pumilio protein binds RNA through a conserved domain that defines a new class of RNA-binding proteins. RNA N. Y. N 3, 1421–1433.