Evolution of structured tandem repeat protein families by exon duplication

Lisanna Paladin[1], Marco Necci[1], Damiano Piovesan[1], Pablo Mier[2], Miguel A. Andrade-Navarro[2], Silvio C.E. Tosatto[1,*]

1 Dept. of Biomedical Sciences, University of Padova, Italy
2 Faculty of Biology, Johannes Gutenberg University of Mainz, Germany
* Corresponding author

Abstract

Tandem Repeat Proteins (TRPs) are ubiquitous in cells, tissues and organisms, and are enriched in eukaryotes. They contributed to the evolution of organism complexity, specializing for functions that require quick adaptability. To investigate the hypothesis of repeat protein evolution through exon duplication and rearrangement, we designed a tool to analyze the relationships between exon/intron patterns and structural symmetries. The tool allows comparison of the structure fragments as defined by exon/intron boundaries from Ensembl against the structural element repetitions from RepeatsDB. The all-against-all pairwise structural alignment between fragments and comparison of the two definitions are visualized in a single matrix, the "repeat/exon plot". An analysis of different repeat protein families, including the solenoids Leucine-Rich, Ankyrin, Pumilio, HEAT repeats and the β propellers Kelch-like, WD40 and RCC1, shows different behaviors, illustrated here through examples. For each example, the analysis of the exon mapping in homologous proteins supports the conservation of their exon patterns. We propose that when a clear-cut relationship between exon and structural boundaries can be identified, it is possible to infer a specific "evolutionary pattern" which may improve TRPs detection and classification.

Abbreviations

TRP	Tandem Repeat Protein
LRR	Leucine-Rich Repeat
ANK	Ankyrin repeat
PUM	Pumilio repeat
HEAT	Huntingtin, elongation factor 3 (EF3), protein phosphatase 2A (PP2A), and the yeast kinase TOR1 repeat
KELCH	Kelch-like repeat
WD40	WD or beta-transducin repeat
RCC1	Regulator of Chromosome Condensation repeat

Introduction

Tandem repeat proteins (TRPs) are a puzzling class of proteins whose 3D architecture consists of the repetition of a simple structural module, called “unit” [1]. Structural units are stabilized by an axis of hydrophobic intra-unit interactions rather than a core [2]. This arrangement confers unique properties to TRPs, including a linear folding pathway where each unit drives the folding of the following [3]. In some cases a binding partner is involved in the stabilization [4]. Their unique arrangement and structural plasticity allow insertions/deletions of units, which can be of remarkable structural diversity provided that they are compatible with the interactions within the stabilizing axis. An additional consequence is that TRPs show a higher surface/volume ratio in comparison to globular proteins. All these features make them a versatile framework for the formation of protein-protein interactions [5]–[7]. TRPs are central in cell signaling and regulation, and widely distributed across functional pathways, performing binding functions that require high evolutionary adaptability [8]. TRPs are abundant across the tree of life, but have specific roles in Eukaryotes and contributed to their evolution [8]–[10]. Their prevalent role as binders and scaffolds was of high importance in the development of eukaryotic signaling and management of complexity and indeed they are far more abundant in multicellular organisms [8]. TRP folds emerged several times across different lineages, arising from the multiple duplication of a segment in coding sequences.

For Eukaryotes, it has been suggested that repeated segments could correspond to exons, thus being easily duplicated and/or shuffled thanks to the modular intron/exon structure [11]. Domains encoded by single exons (exon-bordering domains) in proteins were demonstrated to be not only more abundant and widespread than those that are not [12], but also to show accelerated evolution [13], demonstrating the effectiveness of this framework. However, this refers mainly to autonomously folding domains, while the case of TRPs is more challenging. When comparing exon structure to repeat units, it is important to consider that units are not autonomously folding as they need their neighbours. As a result, an exon coding for a repeat unit is not as evolutionarily versatile as one coding for a full domain as it requires to be shuffled or duplicated in tandem with other repeats. On the other hand, duplication of an exon that includes multiple units is consistent with that requirement, resulting in complex exon patterns. Thus, TRPs are compatible with evolution through intron-facilitated exon duplication, which could also confer them the advantage of easy insertion and removal of units to adapt to different binders. Considering the relationship between TRP sequence and structures [14], this is particularly relevant for TRP families that have a highly variable number of units, either across different organisms or by differential exon expression in cells or conditions. Evidence of exon duplication in proteins has been previously assessed by comparison of exon length and exon/intron phases [15] as well as by alignment between encoded sequences [16]. These features have also been investigated in TRPs, showing correspondence between unit and exon patterns in some TRP single-case studies [17]–[19], but the literature in the field has so far struggled to draw unique conclusions for all repeat folds [20], [21]. One of the most recent studies on the topic establishes that boundaries of structured domains tend to fall in correspondence to exon boundaries, while disordered regions do not [22]. Repeat domains are somehow in-between these two categories, due to their linear and large-surface structure and peculiar folding pathway.

In order to visualize the exon/unit patterns in TRPs, we exploited the RepeatsDB [23], [24] database of tandem repeated protein structures. TRP structures are classified in the database according to the unit length and type of contacts between units. We focused on structures with variable number of units and mainly stabilized by intra-unit interactions, falling into classes III (elongated repeats including solenoids) and IV (closed repeats or toroids). The exon structure of these repeat regions was compared to the structural repeat modularity, to identify patterns of association. We mapped information from RepeatsDB [24], together with structure (PDB [25]) and sequence data (UniProt [26] and Ensembl [27]) and designed a matrix, the repeat-exon plot, that merges useful information to support this comparison: (i) the length and position of exon boundaries and structural units along the protein, (ii) the structural similarity between units, and (iii) the structural similarity between exon-bordering fragments. The complete dataset and code is available at gitlab.com/refract-rise/repeat-exon, featuring information from all sources of data.

Information about the exon/unit relationships in TRPs would be of use to derive their evolutionary mechanisms in relation to structural properties and folding pathways, as well as to support their detection and annotation. In particular, when TRP families show a very consistent exon/units pattern this information provides an evolutionarily related periodicity that should be taken into account in the annotation of repeats in sequences and structures. This is relevant in the context of an ongoing collaboration with Pfam database of protein families [28] to shed light on the non-annotated fraction of proteomes, enriched in disorder and repeats [29].

Materials and Methods

Repeat-exon plot

The repeat-exon plot is a matrix to visualize the pairwise comparison of structural fragments, as defined by exon boundaries and unit boundaries. The matrix shows in the higher half (from top to bottom and left to right) the exon array and on the lower half (from left to right and top to bottom) the unit array along the UniProt sequence length. The matrix is actually a combination of two matrices aligned along the diagonal based on the protein sequence. Rows and columns correspond to structural fragments (either exon- or unit-bordering) and their size is proportional to the fragment size in residues; the background color intensity of a cell is proportional to the TM-score [30] of the pairwise structural alignment between the two corresponding fragments, respectively. High TM-scores are represented by cells of darker color. Exons and structural fragments often do not coincide; the matrix is intended to capture and highlight this shift, apparent when comparing cell sizes along the diagonal (**Figure 1A**).

Although the exon data covers the full-length sequence, repeat units may not do so for two reasons: (i) the structure contains non repeated regions, (ii) the PDB chain is a fragment, i.e. does not entirely cover the UniProt sequence. These non-repeated or missing residues are represented in the units matrix as blank cells. The color pattern of the units and exon matrices pinpoints contiguous exons that code for the same structure, or similarities between units that are non-contiguous. This is illustrated with examples in the results section.

Dataset

Two different types of structural fragments are defined. One corresponds to the exon organization in the sequence and the other is derived by manually annotating the tandem repeated elements in the structure, i.e. the units. The exon definition of the fragments is obtained by combining structure (PDB [25]) and sequence information (UniProt [26], Ensembl [27]) from the corresponding databases. Protein structures from the PDB are mapped to the protein sequences in UniProt entries at the residue level, and these are, in turn, mapped to genes as described in the Ensembl database, which features exon start/end annotation. The alignment between structures, sequences and gene translated sequences allows to match the correct isoform when multiple Ensembl transcript IDs are provided, since the PDB/UniProt mapping provided by SIFTS [31] is not always isoform-specific. The structural repeat definition is obtained from RepeatsDB [24], providing unit start/end positions in TR domains of PDB structures from RepeatsDB-lite [32]. Manually curated entries were selected for the discussion of results and related statistics. RepeatsDB definition of “units” refers to the minimal length repeat that is identifiable in the pattern of structural symmetry. RepeatsDB version 2020.02.18 contains 6,290 entries (PDB chains) mapping to 1,062 UniProt entries of which 755 are eukaryotic. The final dataset comprises 487 proteins which feature Ensembl transcript annotations, i.e. exon mapping. The complete dataset and code is available at gitlab.com/refract-rise/repeat-exon, featuring information from all sources of data.

Statistical analysis

All dataset entries were mapped to UniRef50 and UniRef90 clusters [33] in order to generalize the observations. Within each cluster, eukaryotic proteins with available Ensembl annotation were retained. The exon pattern of each UniRef entry was compared with the patterns of all the others. For each pairwise comparison we calculated an overlap score. To do so, we first established a 1:1 mapping of each exon in the first protein to one exon of the second, based on maximum overlap. Exons that do not pair with any other contribute negatively to the score, as all their corresponding bases are counted as non-overlapping. The final score is obtained by dividing the length (in residues) of overlapping fragments to the total length, and ranges from 0 to 1, where 1 represents the maximum overlap. The scoring system is illustrated in **Figure 1B**. The average of all pairwise comparisons within a UniRef cluster estimates the exon pattern conservation within the cluster. The average of scores of all UniRef90 clusters for the proteins in the dataset is 0.96, indicating that, in general, the exon patterns are conserved across clusters. The exon pattern conservation score was used on a case-by-case basis to generalise the results of the example proteins to their homologs.

Results

To study the correspondence between the periodicity of tandem repeat proteins (TRPs) and their coding exons, we contrasted information of protein structures and exon boundaries mapped to protein sequences of TRPs. Exons and protein domain boundaries have previously been extensively compared to derive information about protein evolution. In this study, we focused on TRP structures in relationship to their exon arrangement, investigating the hypothesis of evolution through exon duplication. We relied on RepeatsDB as a precise source of definitions of TRP units. In RepeatsDB, the definition of the repeated fragments is purely based on structural symmetries, which can be identified via software or by visual inspection. We compared the position of the repeat units with the exon boundaries and evaluated the structural similarity (TM-score) between fragments corresponding to the two different definitions. We designed the repeat/exon matrix to visualize this hypothetical alignment, providing insights into the relationship between units and exons in the repeat region.

We identified different repeat/exon patterns and discussed them through a few illustrative examples conserved in evolutionarily related proteins. The different types of patterns that emerged in the matrices are schematized in **Figure 1C**. In the following paragraphs we describe ten different examples of families with variable number repeats and divided in two groups: solenoids and toroids. The scores of the exon-pattern conservation of the families discussed are summarized in **Table 1**, showing that the exon conservation at 90% of sequence identity is very high and remains well conserved even at 50% of sequence identity.

Solenoids: “scattered”, “checkerboard” and “framed” patterns

Solenoids are arranged in a super-helical fold composed by structurally inter-dependent modules connected by a regular pattern of interactions lacking long range stabilizing contacts [3]. Each unit is composed of a small number of secondary structure elements. Different combinations of α helices and β strands determine the properties of the whole fold, like its flexibility, curvature and twist. Despite their structural similarities, different solenoid families emerged multiple times during evolution [20] and show a variety of exon/unit patterns. In some cases, a single exon spans the whole repeat region, as in the majority of tetratricopeptide repeats (TPRs). In other cases, the repeat region corresponds to several small exons, often showing complex patterns.

Leucine-Rich Repeats (LRRs). Leucine-rich repeat families are generally found in α/β arrangements with different shapes, curvature and rigidity [5], [34]. We identified two main structural sub-classes within LRR regions, namely α/β and β. The two are distinguished by the inter-strand segment, which is always in α helical arrangement in α/β regions (in all the units) while it is not in β regions. α/β LRRs usually show high curvature of the whole region, with internal parallel β sheets and external α helices. Their structure is more regular than β LRRs, which show less curvature (in some cases the units are parallel). The two different arrangements correspond to different patterns in the exon/unit matrices (**Fig.2**). The murine Ribonuclease Inhibitor (**Figure 2A**, PDB 3TSR chain E, UniProtKB Q91VI7) is an α/β LRR in which all exons span two units but the N- and C-terminal ones. The figure shows the full-length protein. All exon boundaries fall exactly in the same position of the β sheet (regular structural phase), and the alignment of structural fragments corresponding to exons have very high scores. Instead, the unit alignments show a “checkerboard” pattern, with each unit aligning better with those not immediately flanking than with those immediately flanking. Together with the regularity of the exon pattern, this suggests that the evolutionary module is the exon comprising two repeat units instead of one. Additional information derived from the matrix is that the N- and C-terminal units (capping units) are different from the others in terms of exon matching, and partially of structure (“framed” pattern). Units with a specific fold that act as a “cap” and stabilize the repeat region are common in TRPs [3], [35], [36]. In this case, our data suggest that they are encoded by specific exons, different from the central ones. β LRRs are flatter and show imperfectly shaped external α helices. The LRR containing G-protein coupled receptor 5 (**Figure 2B**, PDB 4BST chain B, UniProt AC: O75473) has a repeat region (ending approximately at half protein length) with sharp exon phase, i.e. starting always at the same position of the β strand. Exons span different unit numbers (i.e. single, double and triple) and the alignment matrix shows a “scattered” pattern of similarities mainly due to this. The exon pattern and unit phase identified by RepeatsDB are different. Outside of the repeat region, a long exon codes for most of the other half of the protein, a not yet structurally characterized transmembrane segment [37].

Ankyrin repeats (ANKs). The Ankyrin domain is one of the most widely characterized TRP types [6], [36]. Within the analysed families, they show the higher exon pattern conservation score (0.99), conserved at 90% and 50% of sequence identity. We observed that exon boundaries in this family usually fall into the intra-helical loop region. In Ankyrin 1 (**Figure 3A**, PDB 1N11 chain A, UniProt AC: P16157), the repeat region predicted by UniProt spans amino acids 44-795. The available repeat region structure covers amino acids 402-827. The entry shows a very regular phase and length of exons, approximately corresponding to the classical Ankyrin unit size (33 residues) [38]. Two exons have double length (about 66 residues) and span two units. Exons of similar length show high structural alignment scores. Units are also very regular; they match the exon phase and align well. The N- and C-terminal units (number 1 and 12) show slightly lower overall scores in the alignment matrix (framed pattern). The exon length is regular and also compatible with the Ankyrin module in the region at the N-terminus of the analysed PDB structure, confirming the sequence-based prediction reported in UniProt. A similar example is the Ankyrin domain of mouse Tankyrase, which contains five Ankyrin domains, each consisting of four ankyrin repeat units connected by four “linker units”. The latter are approximately 25 residues long, consist of a long and a short α helix and provide a specific curvature angle to the unit array [39]. The structure of Tankirin residues 308-655 (**Figure 3B**, PDB 3UTM chain A, UniProt AC: Q6PFX9) contains three “linker units” and two repeat domains. The unit similarity matrix highlights two areas of high similarity, corresponding to the two domains of three units each, distinguished from the linker areas. Exons match single units inside the Ankyrin domains, with their boundaries falling into the loops between α helices. The last unit inside the repeat domain and the first half of the “linker unit” are instead encoded by a double-unit exon in both cases, and the two longer exons align well. In this case the matrix could be defined as two consecutive framed patterns with inter- and intra-similarities.

Pumilio repeats (PUMs). Pumilio repeats are conserved mRNA binders that can be found in all eukaryotes [40], folding into a solenoidal array of triangular units formed each by two long and one short α helix. They show high promiscuity of interaction partners, including some non-canonical nucleotides [41]. Pum1 (**Figure 3C**, PDB 3BSB chain B, UniProtKB Q14671) and Pum2 were named after their Pumilio repeats [40]. Their repeat domains are encoded by exons of regular size (41/45 residues) corresponding to three α helices (a full triangle) and eventually including a small segment of the following one. The longest among Pum1 TR exons (the structure maps to residues 828-1170) includes a region annotated by RepeatsDB as an insertion. These are structural inclusions within or between repeat units that do not contribute to the stability of the repeat array as they usually have a functional role, e.g. interaction with specific binders. Exon boundaries identify a phase that is different from RepeatsDB annotation (as highlighted by the matrix). Units in RepeatsDB entry 3bsbB are indeed annotated as starting between the two longer helices. The phase of the exons (α helices long-long-short) does therefore not match the phase of the repeat structure (α helices long-short-long).

HEAT repeats. HEAT repeats form solenoidal domains composed of 50 amino acids long units of two α helices linked by a short loop [42]. They are here discussed as an example of a TRP without any clear pattern emerging from the exon/unit matrix analysis. Although the exons are relatively short, they do not have a clear periodicity in the Exportin-1 TR domain (**Figure 3D**, PDB 4BSM chain A, UniProtKB O14980). Structural alignment scores in both the exon and unit matrices are generally lower than in the matrices of the other examples. Of relevance, in the Pfam protein family database [28] HEAT repeats fall in the same clan (CL0020) as tetratricopeptide repeat regions [43]. As already mentioned, the latter do not show a duplication-related exon pattern and are instead usually encoded by a single exon. Pfam clan

CL0020 also includes Armadillo repeats [44] which do not show a clear exon phase either. Not all repeat protein families therefore show a regular exon pattern supporting evolution through exon duplication.

Closed structures: the strange case of the β-propeller

Closed TRPs show a toroidal conformation with the first unit contacting the last. All units are necessary to preserve fold stability and unit insertions are less tolerated than in elongated repeats [1]. Closed structures such as TIM-barrels, β barrels and β trefoils do not show a regular exon/unit pattern and are usually encoded by a single or few exons. However, there is an important exception: β and α/β propellers. This ubiquitous fold is widely used as structural scaffold in eukaryotic organisms and is the result of fold convergence [45]. The various propeller families show different numbers of blades (propeller units), blade orientation and arrangements. Propellers are characterized by a remarkable plasticity compared to other closed structures. In addition, their blades were suggested to have been used as ancestral peptides in protein evolution [46].

The β propeller families show evidence of this plasticity in their exon arrangements. Two main frames of propeller blades were previously identified from the structural point of view. One corresponds to the entire structural blade and the second ("velcro" blade) is characterized by strand-swapping between units, with the most external strand of one unit belonging to the previous one. The "velcro" model stabilizes the closed arrangement of these structures, and evolved from a permutation of the ancestral blade [47]. β propeller exons span in different ways into propeller blades, corresponding to a single blade or the "velcro" ones, including some intermediates. Kelch-like protein 2 (**Figure 4A**, PDB 2XN4 chain A, UniProt AC: O95198) shows a propeller region (residues 294-591) consisting of six copies of the domain Kelch_1 (PF01344). Four single-blade units on the structure map well to their exons, while the other two (second and third) are encoded by two exons: one in the "Velcro" conformation covering most of the second unit and part of the third one, and a shorter exon covering the end of the third unit. DNA damage-binding protein 2 (**Figure 4B**, PDB 3EI2 chain B, UniProt AC: Q2YDS1) includes one propeller domain (the structure maps to residues 60-423) of seven blades. All exons but the third in the repeat region map to almost the entire blade and a small fragment of the following, defining an alternative phase for this β propeller domain annotation. The WD40 propeller region (residues 81-499) of Cell division cycle protein 20 homolog (**Figure 4C**, PDB 4GGA chain A, UniProt AC: Q12834) contains exons spanning approximately one blade, except one spanning almost two units. The α/β propeller in human Regulator of chromosome condensation (**Figure 4D**, PDB 1A12 chain A, UniProt AC: P18754) includes exons with different lengths but almost all of them encode for half a blade and half the following one, once again providing evidence for an evolutionary phase different from the one typically annotated in RepeatsDB.

Discussion

The present study provides a visualization tool to assess the correspondence between exons and structural symmetries in TRPs. The matching between exon and repeat unit patterns observed in some single-case studies supports the hypothesis of repeat evolution through exon duplication and rearrangement [17]–[19], [48]. In order to validate this hypothesis at large, we designed a repeat/exon plot which allows us to recognize the relationship between the periodicities and phases of structural repeat units and exons at first sight. The exon pattern conversation score (see **Table 1**) suggests that the conclusion for the single examples can be generalised to their homologs.

Starting from meaningful examples in RepeatsDB [24], we defined a limited set of possible patterns. The examples discussed above focus on repeats with a variable number of units, in particular solenoids and β propellers. Solenoid data shows that the exon duplication hypothesis is not supported by the repeat/exon pattern for all types of TRPs, e.g. it is not clearly derivable in TPR and HEAT repeats. Interestingly, among the analyzed families, HEAT repeats are also the ones with the lowest exon pattern conservation score in UniRef50, highly decreasing with respect to the UniRef90 values. Instead, we observed a very high regularity of exon/unit patterns in LRR, ANK and PUM domains. In LRRs, we identified two different types of patterns, one with 2:1 and one with 1:1 mapping between structural units and exons, corresponding respectively to α/β and β LRRs. In addition, in α/β LLRs (where usually the exon encodes for two flanking units), we observed a framed pattern with the repeat region terminal units being encoded by single exons. The full picture provides evolutionary clues on the mechanisms of stabilization and diversification of these repeat regions, for example that α/β LRR units may be more stable in pairs, while β LRRs may be more tolerant to unit insertion/deletion. The patterns in ANKs are more scattered, mostly due to the fact that they show exons encoding for both single and double units, with the latter possibly the result of exon fusion. They show high exon pattern conservation (i.e. highest UniRef50 score in **Table 1**) and their structural phase of exon bordering fragments is quite regular and similar to RepeatsDB annotation, with exons usually ending in the loop following two parallel α helices. The same consideration applies to PUMs, where the RepeatsDB annotation phase is however shifted compared to the exon bordering fragments (usually ending in the middle of an α helix). Exon patterns may be meaningful even in cases where no clear relationship with the structural pattern is identifiable, such as HEAT repeats. As discussed above, the HEAT domain is encoded by several short exons of similar size.

In terms of closed repeat structures (i.e. toroids), we focused on β propellers for two reasons: First, they are a heterogeneous repeat subclass in terms of unit numbers and evolutionary history [45]. Probably related to this, β propellers show complex repeat/exon patterns unlike other toroids, which are usually encoded by one or a few exons. Some exons encode for single β propeller blades ending in loops. Others exons encode for so-called “velcro” units (i.e. span parts of two blades) which contribute to β propeller stabilization and usually end within a β sheet. In the last α/β propeller example, all exons correspond to “velcro” units. The β propeller repeat/exon patterns observed support the “velcro” annotation of the structural phase.

In the cases where exon bordering fragments identify a pattern of structural modularity, this “evolutionary phase” can be used as additional input feature for manual annotation and for the definition of templates in template-based repeat detection methods, including both sequence- and structure-based ones.

Conclusion

Tandem repeats in eukaryotes perform unique functions requiring quick adaptability. It has been suggested that some repeat families evolved by exon duplication and rearrangement. Here we designed a repeat/exon matrix to visualize the relationship between exon and structural symmetries. We discussed Leucine Rich, Ankyrin, Pumilio and β propeller repeats where very specific repeat/exon patterns are well conserved inside the same protein family. We facilitated the contextual use of exon information to be used as input feature for both sequence- and structure-based prediction methods, respectively like Pfam [28] and RepeatsDB-lite [32]. Moreover, we make it possible to extend the analysis of repeat/exon patterns inside entire TPR classes to formulate novel evolutionary hypotheses. The repeat/exon plot will be integrated in the RepeatsDB entry page to extend the available information and guide future TRP curation.

Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 823886.

Conflict of Interest: none declared.

Bibliography

[1] A. V. Kajava, “Tandem repeats in proteins: from sequence to structure,” *J. Struct. Biol.*, vol. 179, no. 3, pp. 279–288, Sep. 2012, doi: 10.1016/j.jsb.2011.08.009.

[2] R. Espada, R. G. Parra, M. J. Sippl, T. Mora, A. M. Walczak, and D. U. Ferreiro, “Repeat proteins challenge the concept of structural domains,” *Biochem. Soc. Trans.*, vol. 43, no. 5, pp. 844–849, Oct. 2015, doi: 10.1042/BST20150083.

[3] B. Kobe and A. V. Kajava, “When protein folding is simplified to protein coiling: The continuum of solenoid protein structures,” *Trends Biochem. Sci.*, vol. 25, no. 10, pp. 509–515, 2000, doi: 10.1016/S0968-0004(00)01667-4.

[4] “Folding cooperativity and allosteric function in the tandem-repeat protein class | Philosophical Transactions of the Royal Society B: Biological Sciences.” https://royalsocietypublishing.org/doi/10.1098/rstb.2017.0188 (accessed Apr. 08, 2020).

[5] B. Kobe and a V. Kajava, “The leucine-rich repeat as a protein recognition motif.,” *Curr. Opin. Struct. Biol.*, vol. 11, no. 6, pp. 725–732, 2001, doi: 10.1016/S0959-440X(01)00266-4.

[6] L. K. Mosavi, T. J. Cammett, D. C. Desrosiers, and Z. Peng, “The ankyrin repeat as molecular architecture for protein recognition,” *Protein Sci. Publ. Protein Soc.*, vol. 13, no. 6, pp. 1435–1448, Jun. 2004, doi: 10.1110/ps.03554604.

[7] T. F. Smith, “Diversity of WD-repeat proteins,” *Subcell. Biochem.*, vol. 48, pp. 20–30, 2008, doi: 10.1007/978-0-387-09595-0_3.

[8] M. A. Andrade, C. Perez-Iratxeta, and C. P. Ponting, “Protein repeats: structures, functions, and evolution,” *J. Struct. Biol.*, vol. 134, no. 2–3, pp. 117–131, Jun. 2001, doi: 10.1006/jsbi.2001.4392.

[9] A. Schüler and E. Bornberg-Bauer, “Evolution of Protein Domain Repeats in Metazoa,” *Mol. Biol. Evol.*, vol. 33, no. 12, pp. 3170–3182, Dec. 2016, doi: 10.1093/molbev/msw194.

[10] E. Schaper and M. Anisimova, “The evolution and function of protein tandem repeats in plants,” *New Phytol.*, vol. 206, no. 1, pp. 397–410, Apr. 2015, doi: 10.1111/nph.13184.

[11] M. Liu and A. Grigoriev, “Protein domains correlate strongly with exons in multiple eukaryotic genomes--evidence of exon shuffling?,” *Trends Genet. TIG*, vol. 20, no. 9, pp. 399–403, Sep. 2004, doi: 10.1016/j.tig.2004.06.013.

[12] M. Liu, S. Wu, H. Walch, and A. Grigoriev, “Exon-domain correlation and its corollaries,” *Bioinforma. Oxf. Engl.*, vol. 21, no. 15, pp. 3213–3216, Aug. 2005, doi: 10.1093/bioinformatics/bti509.

[13] B. Lorente-Galdos *et al.*, “Accelerated exon evolution within primate segmental duplications,” *Genome Biol.*, vol. 14, no. 1, p. R9, 2013, doi: 10.1186/gb-2013-14-1-r9.

[14] L. Paladin and S. C. E. Tosatto, “Comparison of protein repeat classifications based on structure and sequence families,” *Biochem. Soc. Trans.*, vol. 43, no. 5, pp. 832–837, Oct.

2015, doi: 10.1042/BST20150079.

[15] A. Fedorov, L. Fedorova, V. Starshenko, V. Filatov, and E. Grigor'ev, "Influence of Exon Duplication on Intron and Exon Phase Distribution," *J. Mol. Evol.*, vol. 46, no. 3, pp. 263–271, Mar. 1998, doi: 10.1007/PL00006302.

[16] A. Broom, K. Trainor, D. W. MacKenzie, and E. M. Meiering, "Using natural sequences and modularity to design common and novel protein topologies," *Curr. Opin. Struct. Biol.*, vol. 38, pp. 26–36, Giugno 2016, doi: 10.1016/j.sbi.2016.05.007.

[17] M. C. Haigis, E. S. Haag, and R. T. Raines, "Evolution of ribonuclease inhibitor by exon duplication," *Mol. Biol. Evol.*, vol. 19, no. 6, pp. 959–963, Jun. 2002, doi: 10.1093/oxfordjournals.molbev.a004153.

[18] A. K. Björklund, S. Light, R. Sagit, and A. Elofsson, "Nebulin: a study of protein repeat evolution," *J. Mol. Biol.*, vol. 402, no. 1, pp. 38–51, Sep. 2010, doi: 10.1016/j.jmb.2010.07.011.

[19] S. Light, R. Sagit, S. S. Ithychanda, J. Qin, and A. Elofsson, "The evolution of filamin-a protein domain repeat perspective," *J. Struct. Biol.*, vol. 179, no. 3, pp. 289–298, Sep. 2012, doi: 10.1016/j.jsb.2012.02.010.

[20] E. Schaper, O. Gascuel, and M. Anisimova, "Deep conservation of human protein tandem repeats within the eukaryotes," *Mol. Biol. Evol.*, vol. 31, no. 5, pp. 1132–1148, May 2014, doi: 10.1093/molbev/msu062.

[21] T. O. Street, G. D. Rose, and D. Barrick, "The Role of Introns in Repeat Protein Gene Formation," *J. Mol. Biol.*, vol. 360, no. 2, pp. 258–266, Jul. 2006, doi: 10.1016/j.jmb.2006.05.024.

[22] B. Smithers, M. Oates, and J. Gough, "'Why genes in pieces?'—revisited," *Nucleic Acids Res.*, doi: 10.1093/nar/gkz284.

[23] T. Di Domenico *et al.*, "RepeatsDB: a database of tandem repeat protein structures," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D352-357, Jan. 2014, doi: 10.1093/nar/gkt1175.

[24] L. Paladin, L. Hirsh, D. Piovesan, M. A. Andrade-Navarro, A. V. Kajava, and S. C. E. Tosatto, "RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D308–D312, 04 2017, doi: 10.1093/nar/gkw1136.

[25] "RCSB Protein Data Bank: Enabling biomedical research and drug discovery - Goodsell - 2020 - Protein Science - Wiley Online Library." https://onlinelibrary.wiley.com/doi/full/10.1002/pro.3730 (accessed Apr. 08, 2020).

[26] UniProt Consortium, "UniProt: a worldwide hub of protein knowledge," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/nar/gky1049.

[27] F. Cunningham *et al.*, "Ensembl 2019," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D745–D751, Jan. 2019, doi: 10.1093/nar/gky1113.

[28] S. El-Gebali *et al.*, "The Pfam protein families database in 2019," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D427–D432, Jan. 2019, doi: 10.1093/nar/gky995.

[29] J. Mistry *et al.*, "The challenge of increasing Pfam coverage of the human proteome," *Database J. Biol. Databases Curation*, vol. 2013, p. bat023, 2013, doi: 10.1093/database/bat023.

[30] Y. Zhang and J. Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," *Nucleic Acids Res.*, vol. 33, no. 7, pp. 2302–2309, 2005, doi: 10.1093/nar/gki524.

[31] J. M. Dana *et al.*, "SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations

for proteins," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D482–D489, Jan. 2019, doi: 10.1093/nar/gky1114.

[32] L. Hirsh, L. Paladin, D. Piovesan, and S. C. E. Tosatto, "RepeatsDB-lite: a web server for unit annotation of tandem repeat proteins," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W402–W407, 02 2018, doi: 10.1093/nar/gky360.

[33] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and UniProt Consortium, "UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches," *Bioinforma. Oxf. Engl.*, vol. 31, no. 6, pp. 926–932, Mar. 2015, doi: 10.1093/bioinformatics/btu739.

[34] J. Bella, K. L. Hindle, P. A. McEwan, and S. C. Lovell, "The leucine-rich repeat structure," *Cell Mol Life Sci*, vol. 65, pp. 2307–2333, 2008.

[35] A. V. Kajava and A. C. Steven, "β-Rolls, β-Helices, and Other β-Solenoid Proteins," in *Advances in Protein Chemistry*, vol. Volume 73, J. M. S. and D. A. D. P. Andrey Kajava, Ed. Academic Press, 2006, pp. 55–96.

[36] Z. Islam, R. S. K. Nagampalli, M. T. Fatima, and G. M. Ashraf, "New paradigm in ankyrin repeats: Beyond protein-protein interaction module," *Int. J. Biol. Macromol.*, vol. 109, pp. 1164–1173, Apr. 2018, doi: 10.1016/j.ijbiomac.2017.11.101.

[37] K. K. Kumar, A. W. Burgess, and J. M. Gulbis, "Structure and function of LGR5: an enigmatic G-protein coupled receptor marking stem cells," *Protein Sci. Publ. Protein Soc.*, vol. 23, no. 5, pp. 551–565, May 2014, doi: 10.1002/pro.2446.

[38] H. K. Binz, M. T. Stumpp, P. Forrer, P. Amstutz, and A. Plückthun, "Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins," *J. Mol. Biol.*, vol. 332, no. 2, pp. 489–503, Sep. 2003, doi: 10.1016/S0022-2836(03)00896-9.

[39] S. Morrone, Z. Cheng, R. T. Moon, F. Cong, and W. Xu, "Crystal structure of a Tankyrase-Axin complex and its implications for Axin turnover and Tankyrase substrate recruitment," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 5, pp. 1500–1505, Jan. 2012, doi: 10.1073/pnas.1116618109.

[40] L. K. Dedow and J. Bailey-Serres, "Searching for a Match: Structure, Function and Application of Sequence-Specific RNA-Binding Proteins," *Plant Cell Physiol.*, vol. 60, no. 9, pp. 1927–1938, Sep. 2019, doi: 10.1093/pcp/pcz072.

[41] Y. K. Gupta, D. T. Nair, R. P. Wharton, and A. K. Aggarwal, "Structures of Human Pumilio with Noncognate RNAs Reveal Molecular Mechanisms for Binding Promiscuity," *Structure*, vol. 16, no. 4, pp. 549–557, Apr. 2008, doi: 10.1016/j.str.2008.01.006.

[42] M. A. Andrade, C. Petosa, S. I. O'Donoghue, C. W. Müller, and P. Bork, "Comparison of ARM and HEAT protein repeats," *J. Mol. Biol.*, vol. 309, no. 1, pp. 1–18, May 2001, doi: 10.1006/jmbi.2001.4624.

[43] A. Perez-Riba and L. S. Itzhaki, "The tetratricopeptide-repeat motif is a versatile platform that enables diverse modes of molecular recognition," *Curr. Opin. Struct. Biol.*, vol. 54, pp. 43–49, Feb. 2019, doi: 10.1016/j.sbi.2018.12.004.

[44] R. Tewari, E. Bailes, K. a Bunting, and J. C. Coates, "Armadillo-repeat protein functions: questions for little creatures," *Trends Cell Biol*, vol. 20, pp. 470–481, 2010.

[45] I. Chaudhuri, J. Söding, and A. N. Lupas, "Evolution of the beta-propeller fold," *Proteins*, vol. 71, no. 2, pp. 795–803, May 2008, doi: 10.1002/prot.21764.

[46] K. O. Kopec and A. N. Lupas, "β-Propeller blades as ancestral peptides in protein evolution," *PloS One*, vol. 8, no. 10, p. e77074, 2013, doi: 10.1371/journal.pone.0077074.

[47] P. L. Clark, "How to Build a Complex, Functional Propeller Protein, From Parts," *Trends Biochem. Sci.*, vol. 41, no. 4, pp. 290–292, Apr. 2016, doi: 10.1016/j.tibs.2016.02.010.

[48] G. Liu *et al.*, “Analysis of Primate Genomic Variation Reveals a Repeat-Driven Expansion of the Human Genome,” *Genome Res.*, vol. 13, no. 3, pp. 358–368, Mar. 2003, doi: 10.1101/gr.923303.

Tables

Family	Example	UniRef cluster	Proteins in cluster	Proteins annotated by Ensembl	Exon pattern score
Solenoids					
LRR	Q91VI7	UniRef90_Q91VI7	5	4	0.95
		UniRef50_P10775	20	8	0.80
LRR	O75473	UniRef90_O75473	42	7	0.89
		UniRef50_O75473	276	125	0.95
ANK	P16157	UniRef90_P16157	2	1	1
		UniRef50_P16157	140	42	0.99
ANK	Q6PFX9	UniRef90_O95271	80	56	0.99
		UniRef50_O95271	431	164	0.97
PUM	Q14671	UniRef90_Q14671	129	48	0.98
		UniRef50_Q14671	312	93	0.91
HEAT	O14980	UniRef90_O14980	35	19	0.93
		UniRef50_O14980	537	117	0.76
β propellers					
KELCH	O95198	UniRef90_O95198	156	94	0.93
		UniRef50_O95198	194	100	0.96
WD40	Q2YDS1	UniRef90_Q2YDS1	4	4	0.95
		UniRef50_Q99J79	56	11	0.93
WD40	Q12834	UniRef90_Q12834	58	27	0.97
		UniRef50_Q12834	238	61	0.85
RCC1	P18754	UniRef90_P18754	47	19	0.72
		UniRef50_P18754	228	91	0.83

Table 1: Statistics about the repeat protein families discussed. For each of the examples, the table reports the UniRef cluster references to UniRef90 and UniRef50, the number of

proteins contained in the clusters, the number of proteins within the cluster annotated with exon position by Ensembl and the exon pattern conservation score derived by this sub-cluster. Underlined, the maximum and minimum exon pattern conservation score value at 50% sequence identity.

Figures

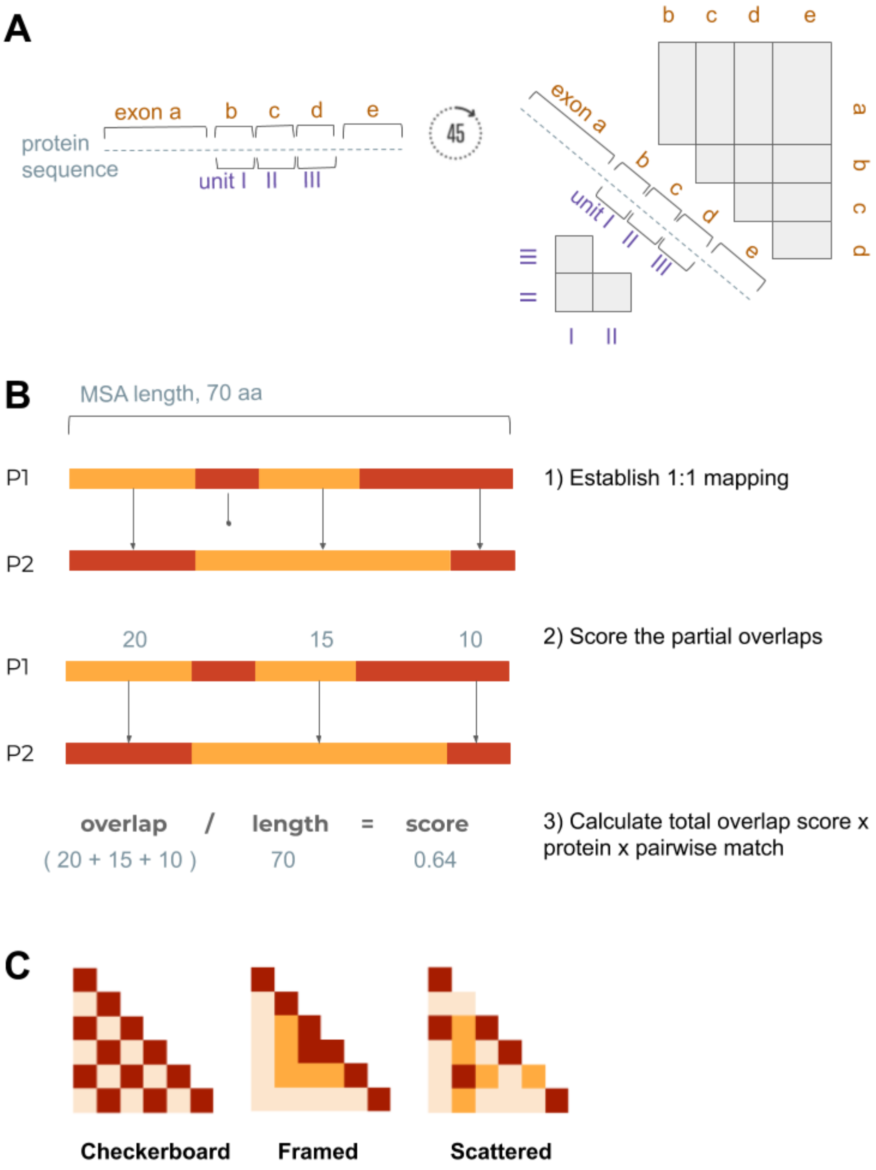

Figure 1: A) Repeat-exon plot structure. Exon and repeat unit positions are mappe over the protein sequence. Repeat units may not cover the entire UniProt sequence, while exons do. The matching between exon and unit periodicity and phase can be observed along the matrix diagonal. The matrix cells (here in grey) are colored according to the structural similarity between fragments. **B) Overlap scoring system.** It is used to score the exon regularity within a UniRef cluster; here simplified with two proteins (P1 and P2) with a multiple sequence alignment of length 70. The lengths of overlapping segments in matching exons is summed, and this total is then divided by the MSA length to obtain the overlap score. **C) Matrix patterns.** Three possible matrix configurations (checkerboard, framed and scattered) corresponding to different evolution patterns.

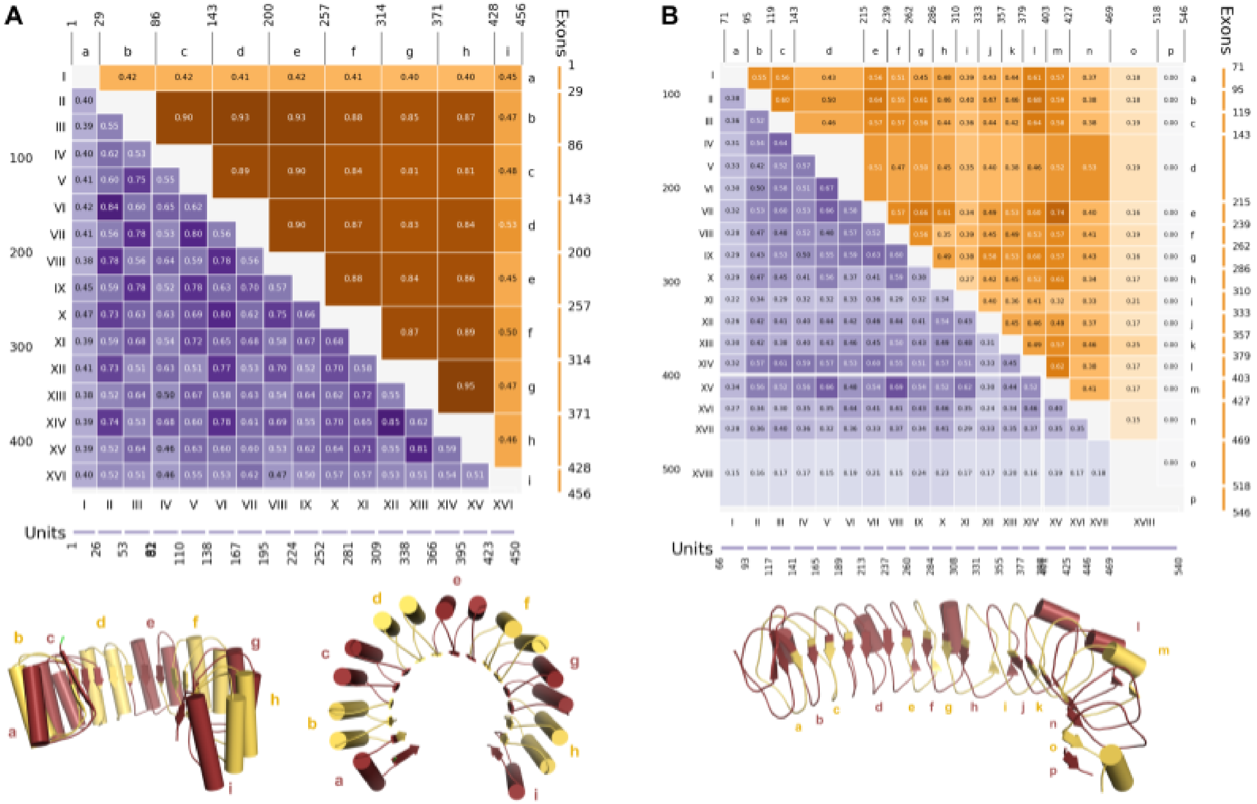

Figure 2: Leucine Rich Repeats matrices. The colors in the two submatrices represent the structural similarities between fragments (either units, in purple, or exons, in orange). Darker colors correspond to higher structural similarity scores. Exons are mapped in alternating colors in the structure below the matrix. The first (on the left) is a full length protein, entirely repeated, including eight exons and sixteen units. Each internal exon (b-h) corresponds to two units, while the terminal ones (a and i) map to one unit only. Exons b-h are periodic and once mapped to the structure (shown in alternating colors in the figure below the matrix) they identify a structural phase that ends in the middle of each second sheet. The second (on the right) shows the repeat region in structure from residue 71 to 546. The domain following is mostly unstructured. **A) PDB 3TSR chain E**, UniProt AC: Q91VI7. **B) PDB 4BST chain B**, UniProt AC: O75473.

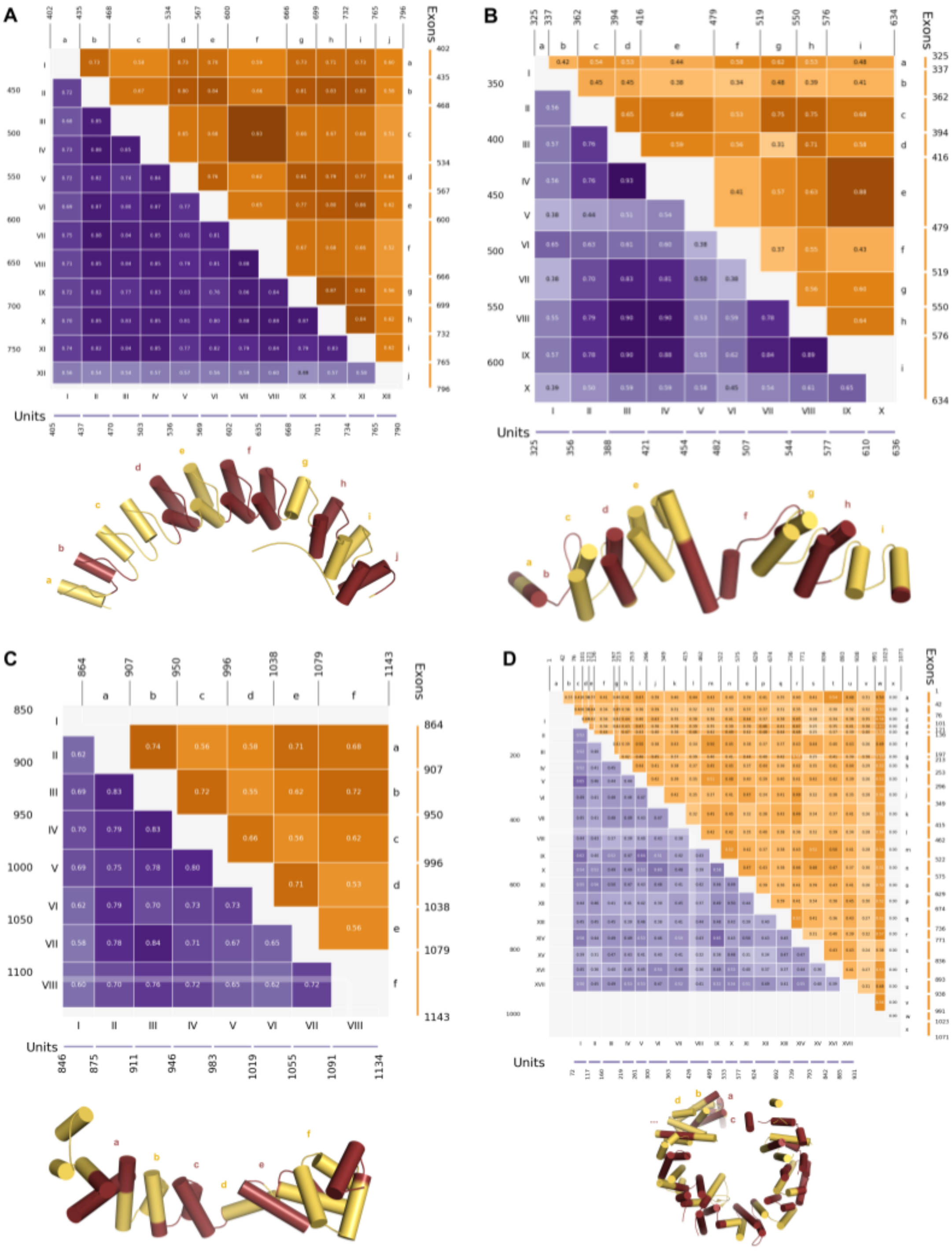

Figure 3: Other solenoids matrices. The colors in the two submatrices represent the structural similarities between fragments (either units, in purple, or exons, in orange). Darker colors correspond to higher structural similarity scores. Exons are mapped in alternating colors in the structure below the matrix. **A) PDB 1N11 chain A**, UniProt AC: P16157. **B) PDB 3UTM chain A**, UniProt AC: Q6PFX9. **C) PDB 3BSB chain B**, UniProt AC: Q14671. **D) PDB 4BSM chain A**, UniProt AC: O14980.

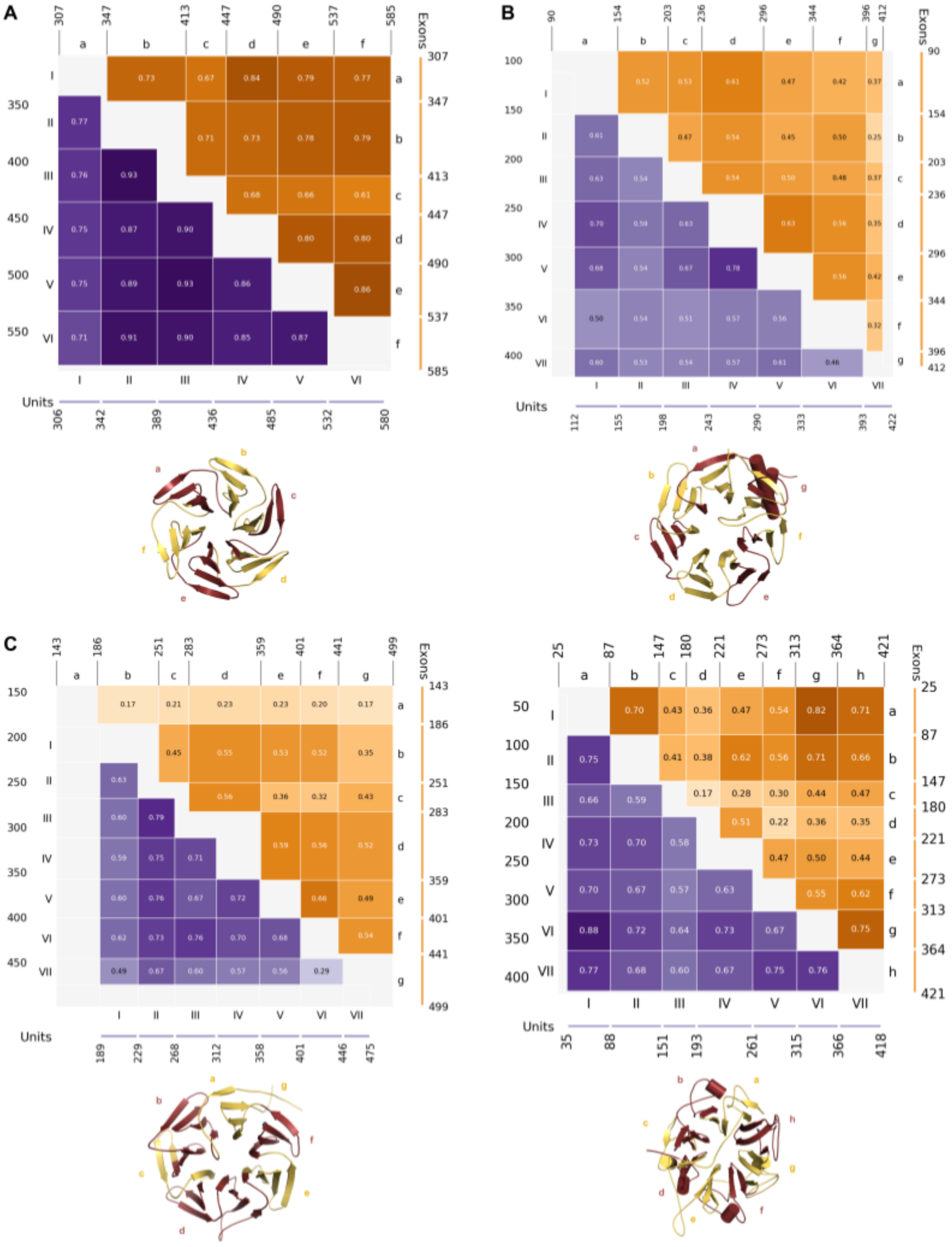

Figure 4: Beta propellers matrices. The colors in the two submatrices represent the structural similarities between fragments (either units, in purple, or exons, in orange). Darker colors correspond to higher structural similarity scores. Exons are mapped in alternating colors in the structure below the matrix. **A) PDB 2XN4 chain A**, UniProt AC: O95198. **B) PDB 3EI2 chain B**, UniProt AC: Q2YDS1. **C) PDB 4GGA chain A**, UniProt AC: Q12834. **D) PDB 1A12 chain A**, UniProt AC: P18754.