Why does the first protein repeat often become the only one?

Simona Manasra, Andrey V Kajava

Why does the first protein repeat often become the only one?

Simona Manasra[1] and Andrey V. Kajava[2]*

[1] Institute of Bioengineering, ITMO University, Kronverksky Pr. 49, 197101 Saint Petersburg, Russia
[2] Centre de Recherche en Biologie cellulaire de Montpellier (CRBM), UMR 5237 CNRS, Université Montpellier, 1919 Route de Mende, Cedex 5, 34293 Montpellier, France

* Corresponding author: E-mail: andrey.kajava@crbm.cnrs.fr

Abstract

Proteins with two similar motifs in tandem are one of the most common cases of tandem repeat proteins. The question arises: why is the first emerged repeat frequently fixed in the process of evolution, despite the ample opportunities to continue its multiplication at the DNA level? To answer this question, we systematically analyzed the structure and function of these proteins. Our analysis showed that, in the vast majority of cases, the structural repetitive units have a two-fold (C2) internal symmetry. These closed structures provide an internal structural limitation for the subsequent growth of the repeat number. Frequently, the units "swap" their secondary structure elements with each other. Moreover, the duplicated domains, in contrast to other tandem repeat proteins, form binding sites for small molecules around the axis of C2 symmetry. Thus, the closure of the C2 structures and the emergence of new functional sites around the axis of C2 symmetry provide plausible explanations for why a repeat, once appeared, becomes fixed in the evolutionary process. We have placed these structures within the general structural classification of tandem repeat proteins, classifying them as either Class IV or V depending on the size of the repetitive unit.

Keywords: protein repeats, symmetry, 3D structure, classification, swapping, evolution

1. Introduction

Proteins contain a large portion of sequences with arrays of repeats that are directly adjacent to each other (Heringa, 1998; Marcotte et al., 1999; Kajava, 2001; Andrade et al., 2001). These tandem repeats (TRs) are ranging from the repetition of a single amino acid to motifs of 150 or more residues. Approximately half of TR regions are naturally unfolded (Tompa, 2003; Simon and Hancock, 2009; Jorda et al., 2010). The other half of them folds in a variety of either elongated or ring-like structures having a continuum of shapes, which contrast sharply with globular shapes of non-repetitive proteins. Despite a great diversity of shapes, the structures of TRs can be well-described in a synthetic manner by using a classification based on the length of their repetitive units (Kajava, 2012). In this classification, the protein structures can be subdivided into five classes: (1) crystalline aggregates formed by repeats of 1 or 2 residues, (2) fibrous oligomers stabilized by inter-chain interactions with 3-7 residue repeats, (3) elongated structures with the repeat length of 5–40 residues such as, for example, solenoid folds, (4) closed (ring-like) structures consisting of 30-60 residues repeats, (5) beads on a string structures with a typical size of repeats over 50 residues, which are large enough to fold

independently into stable domains. The non-globular 3D structures of TR proteins with the unusually large solvent-accessible surface area allow them to carry out specific biological functions forming, for example, building blocks of supramolecular structures, binders of periodic surfaces and molecules, or bind ligands of different types and sizes (Andrade et al., 2001; Kajava, 2012, Paladin et al.,, 2021).

From an evolutionary standpoint, TRs have several specific features. Extension or shortening of the repeating sequences is known to be the result of unequal crossing-over between TR sequences during meiosis occurring between misaligned chromatids on homologous chromosomes (Buard and Vergnaud, 1994; Andrade et al., 2001). This error-prone process occurs far more frequently than the background rate of point mutations and represents the basis for quicker evolution of the repetitive sequences in comparison with nonrepetitive ones. As a result, the number of repeats can vary even between orthologues (Saupe et al., 1995). Selective functional advantage of multiple repeats results in these TRs being fixed among populations.

The critical event for the initiation of TR sequences is the first intragenic duplication, which sets the stage for unequal crossing-over and further functional specialization (Andrade et al., 2001). The emergence of the first repeat is a rare event in comparison with the multiplication/reduction of already existing TRs. Its mechanism is still poorly understood. At the same time, proteins with only two domains in tandem are one of the most common cases representing over 20% among TR proteins with the known 3D structure (see Results). The question arises, why the first emerged repeat, being able to either multiply or regress to a single copy domain, nevertheless often becomes fixed in the process of evolution. Since this phenomenon cannot be explained at the DNA level, in this work, we systematically analyzed the structure and function of the proteins with only two repetitive units in order to answer this question. It is important to note that this significant class of proteins has not yet received sufficient attention from the scientific community. The primary focus of researchers has been on proteins containing three or more repeats, and the majority of bioinformatics tools, analyses, and classification methods have been developed for these multi-repeat proteins (Kajava, 2001; Paladin et al., 2021). To fill this gap, we developed bioinformatics protocols to select the known 3D structures with only 2 repetitive units in tandem (abbreviated here as 2RUT) and to describe their symmetries. This allowed us to classify these structures and described their typical functions. The results of this work shed light on why the first repeat frequently may become evolutionarily conserved.

2. Methods

2.1 Selection of protein structures containing 2RUT.

We used the PDB Release of March 2021 for the analysis (Berman et al., 2000). Since the PDB contains the 3D structures of many identical or similar proteins, it was reduced using the CD-HIT program (Huang et al., 2010) to have only one representative structure for sequences with more than 80% identity. Then, we removed from this dataset structures consisting of less than 60 amino acid residues assuming that most of the repeats with less than 30 amino acid residues are likely to have only one element of the secondary structure, for example, an α-helix. As a result, our non-redundant set had 46595 protein structures. To select a set of proteins with two similar structural motifs arranged in tandem, we ran TAPO program (Do Viet et al., 2015) against these 46595 protein structures. The TAPO program searches for repetitive structural motifs in a protein by calculating different features such as RMSDs of its fragments, by the periodicity of strings generated by conformational alphabets, residue contact maps, and arrangements of vectors of secondary structure elements. Then, the dataset of proteins with

2RUTs selected by TAPO was reduced by the CD-HIT to have one representative structure for sequences with more than 60% identity.
TAPO found 4717 protein structures with 2RUT. Further, these structures were annotated using Pfam and CATH databases (Mistry et al., 2021; Sillitoe et al., 2021) and only one representative structure was left if several of them had the same CATH and Pfam ID. The detection of Pfams was made using HMMER search (http://hmmer.org) against the Pfam database. Some of the selected proteins had three and more repeats in tandem of the complete sequences but their 3D structures in PDB had two units because of the shorter fragments used for crystallization or NMR studies. Therefore, we also checked if the proteins with a given Pfam ID have domain architectures with more than two repetitive motifs. For this purpose, we calculated a parameter «Pfam2ratio»*, which is the ratio between the number of times the domain occurs in the sequence more than twice in a row and the number of all sequences where the domain occurs. Then, we manually examined the structures with values of «Pfam2ratio» more than 10% and removed ones, which contained more than two repetitive units. In addition, our manual analysis of several randomly selected structures showed that a number of them have Rossmann fold with several α/β units. They were erroneously included in our dataset because TAPO detected only two α/β units from several ones in their structures. Therefore, we deleted the structures defined as Rossmann Fold by CATH, which had repeat unit length < 85 residues and repeat cover < 0.5. After the application of all mentioned filters, we obtained 1198 unique protein structures containing 2RUT (Supplementary Data 1).

2.2 Methods used to obtain characteristics of the selected structures.
The identified structures have the following information generated by TAPO: PDB code, ID of the chain, positions of the beginning and end of each 3D repeat, the scores evaluated the probability of correct detection of the repeats. We also obtained the secondary structure of the repeats by using the DSSP algorithm from the Bio.PDB module (DSSP of the Biopython package).
Internal symmetry of the structures was determined by using CE-Symm program (Bliven et al., 2019). In addition, the C2 symmetry was detected by our own algorithm described in the section 3.3. Detection of swapping between the repetitive elements was made by using an algorithm developed in this work (see section 3.5). As a result, all structures from our dataset were characterized by the length of the repeats, secondary structure, fraction of the chain occupied by the repeat, domain classification from CATH and Pfam databases, frequency of TR to have more than two Pfams, type of the internal symmetry, the presence/absence of swapping (Supplementary Data 1). These parameters were further used to classify and cluster structures by the k-means method.

2.3 Clustering of the structures
Clustering was carried out separately for proteins with repeat lengths of less than and more than 55 amino acid residues because the 2RUTs with repeat lengths of less than 55 residues frequently form one common structural domain while the ones of more than 55 residues are able to form two structural domains. The clustering was made by using the KMeans module of the scikit-learn library for the Python programming language (Pedregosa et al., 2011). The method is based on minimizing the total square deviation of cluster points in n-dimensional (n features) space from cluster centers. The number of clusters was chosen based on the dependence of inertia on the possible number of clusters calculated using the KMeans module. Inertia, also called the within cluster sum of squares, is the principal measure of clustering performance, which measures how far apart the points are in each cluster. We also clustered the structures with Foldseek all-against-all structural alignment method (Kempen et al., 2022),

which is making a sequence from a structure with 3D-interactions (3Di) alphabet for the further alignment.

3. Results and discussion

3.1 Distribution and characteristics of 2RUTs in proteins

To estimate the prevalence of the proteins with 2RUT, first, we analyzed a non-redundant protein set (sequence identity lower than 80%) from the PDB by a pipeline MetaRepeatFinder (MRF) (https://bioinfo.
crbm.cnrs.fr/index.php?route=tools&tool=15 (Richard and Kajava, 2014)) developed to detect TRs in protein sequence. Proteins with the repeat length of more than 30 residues were considered. The analysis showed that the cases with 2RUTs are one of the most occurring and second only to TRs with three units (Fig 1A). It is known that the PDB has a bias towards small globular structures amenable for the crystallization and NMR studies in contrast to the large elongated structures of TR-containing proteins. Hence, for a more accurate estimation of the prevalence of 2RUT proteins, it is advisable to employ a protein set derived from well-established proteomes. In line with this objective, we further analyzed the human proteome using MRF. The analysis revealed a comparable distribution in the occurrence of proteins with varying numbers of repeats (Fig S1 in Supplementary Data 2). It is known that structure-based methods for detecting tandem repeats generally exhibit higher sensitivity compared to sequence-based methods alone. Therefore, to compare the occurrence of 2RUT structures with those that contain more repetitive elements, we also used the results of the TAPO run against the same non-redundant set from the PDB (Fig 1B). The analysis revealed that the 2RUT structures vastly over numbers the other repetitive structures. This once again proved the importance of studying these structures.

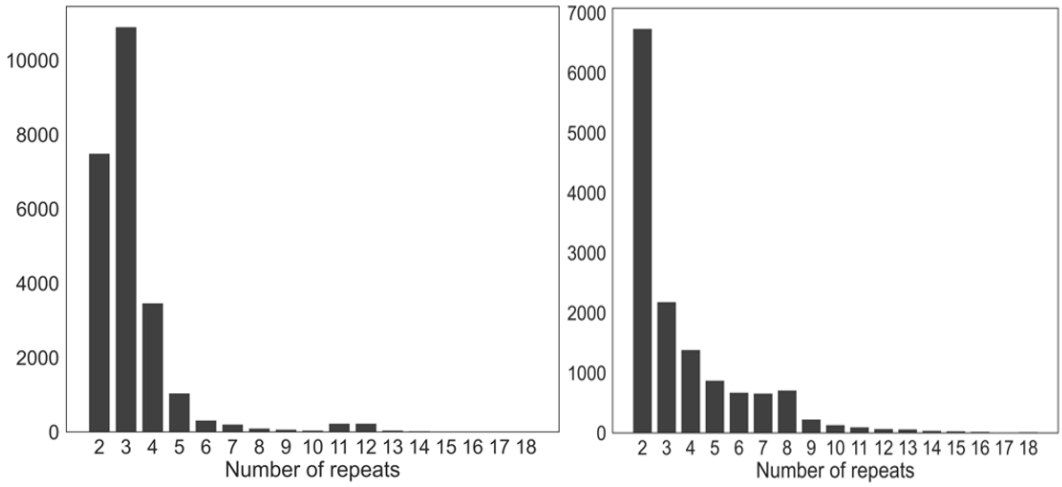

Figure 1. (A) Occurrence of tandem repeats with a different number of repeats in the non-redundant set of proteins from PDB. The repeats were identified by a pipeline MetaRepeatFinder (MRF) (https://bioinfo.crbm.cnrs.fr/index.php?route=tools&tool=15 (Richard and Kajava, 2014)), and only proteins with the repeat length of more than 30 residues were considered. To avoid redundancy, proteins having a level of identity higher than 80% were clustered by CD-HIT (Y. Huang et al., 2010) and represented by only one protein. (B) Occurrence of TRs found in the non-redundant set of the PDB

structures by TAPO. Only structures having the repeat length of more than 30 residues were considered. The results show that 2RUT proteins represent 48% of all TRs.

Following the protocol described in Methods, we obtained 1198 structures containing two repetitive units (Supplementary Data 1) with 60% sequence similarity cutoff and unique PFAM and CATH entries. Our survey of these structures suggests that, in general, they can be subdivided into two groups. The structures with the repeat length less than approximately 55 residues fold into one structured domain, and ones with longer repeats usually have two domains each corresponding to one repeat. The analysis of the occurrence of the structures depending on the repeat length shows that 2RUT structures are especially frequent in the range of repeats between 30 and 55 residues (Fig 2). 2RUT structures containing repeats of over 200 residues are rare. These structures may contain several structural domains in one repetitive unit.

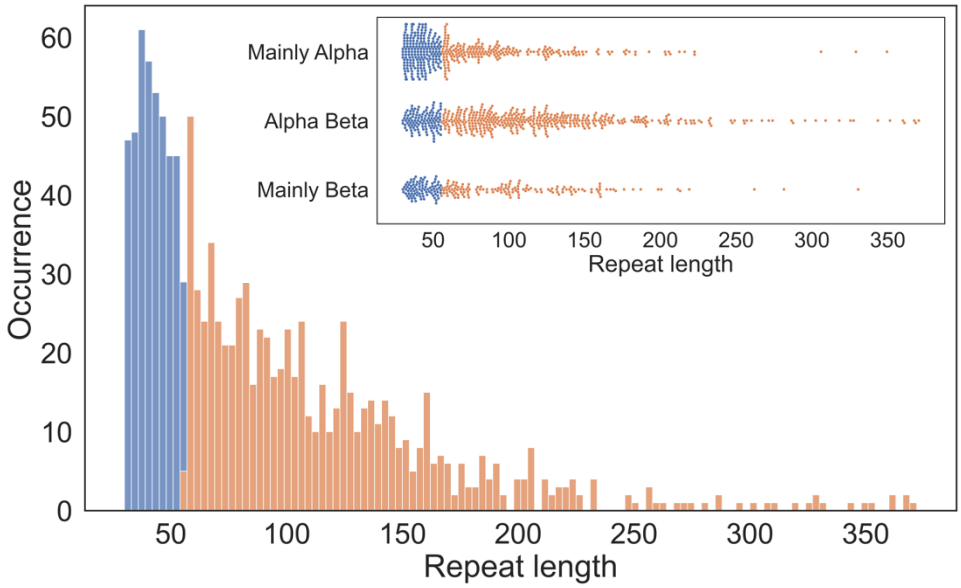

Figure 2. Occurrence of structures with 2RUT depending on the repeat length. The repeat length starts from 30 residues. Structures with repeat lengths ranging between 30 and 55 are in blue, while structures with repeat lengths exceeding 55 are represented in orange. An insert with swarm plots shows the distribution of secondary structures depending on the repeat length.

The analysis of the secondary structures shows the prevalence of α-helices in the structures with repeats less than 55 residues (**α=52.0%**; α/β=26.7%; β=21.0%) and α/β structures (α=28.1%; **α/β=53.6%**; β=18.3%) in cases with repeats more than 55 residues (an insert in Fig. 2). This distinction is due to the frequent presence of four-helix bundles in the 2RUT structures with repeats of less than 55 residues.

3.2 Internal symmetry of 2RUT structures

Protein structures built of repetitive structural units frequently have an internal symmetric arrangement, which is associated with a specific biological function and allows us to trace back evolutionary events linked to the formation of this structure (Goodsell and Olson, 2000; Bliven et al., 2019). The repetitive structures can be subdivided into ones having closed and open symmetries. The closed symmetries of proteins can be further classified as cyclic (C*n*),

generated by a single *n*-fold rotational operator; dihedral (D*n*), which requires an *n*-fold rotation and *n*-perpendicular 2-fold operators (Bliven et al., 2019). Typically, repetitive structures with closed symmetry have a strict limit on the number of repeats. Structures with the most common open symmetries have helical (H), translational (R), and superhelical (SH) arrangements of the repetitive units. In principle, these structures can have unlimited number of repeats.

Since the structures with one repeat have two repetitive units it was interesting to know whether they are arranged symmetrically or not and if yes, what kind of symmetry they have. Several symmetry detection algorithms have been developed (reviewed in Myers-Turnbull et al., 2014). In this work, we used CE-Symm program (version 2.22), which was tested and performed well on the benchmark sets and can be downloaded locally for large-scale analyses (Myers-Turnbull et al., 2014; Bliven et al., 2019). The CE-Symm was run against our set of 2RUT proteins analyzing only those parts of the structures that contain two repetitive units previously detected by TAPO. Among 1198 protein structures, we identified 583 with C1, 528 with C2, 70 with R, 15 with D2, 1 with H, and 1 with C3 symmetry (Supplementary Data 1). Our manual analysis revealed a larger number of structures containing C2 symmetry, which were not detected by CE-Symm program. Therefore, we complemented the CE-Symm results with ones obtained by a new approach developed in this work (see the next section).

3.3 Method for detection of C2 symmetry based on contacts between two repetitive units.

In the C2 arrangement, inter-repeat contacts between structurally equivalent positions exist in pairs n-m’ and n’-m (Fig 3), where n, m and n’, m’ are residues at n- and m-positions of the first and second repetitive unit, respectively. Since the repetitive units are often not identical in terms of sequence and structure, the equivalence of the residue positions was assessed using structural superposition of two domains. Our algorithm was based on the analysis of inter-unit contacts and if the analyzed structure had at least one pair n-m’; n’-m, it was considered to have C2 symmetry. The benchmark of this method showed that the best detection of C2 was obtained when we allowed an error ±3 in the assignment of equivalent residue positions.

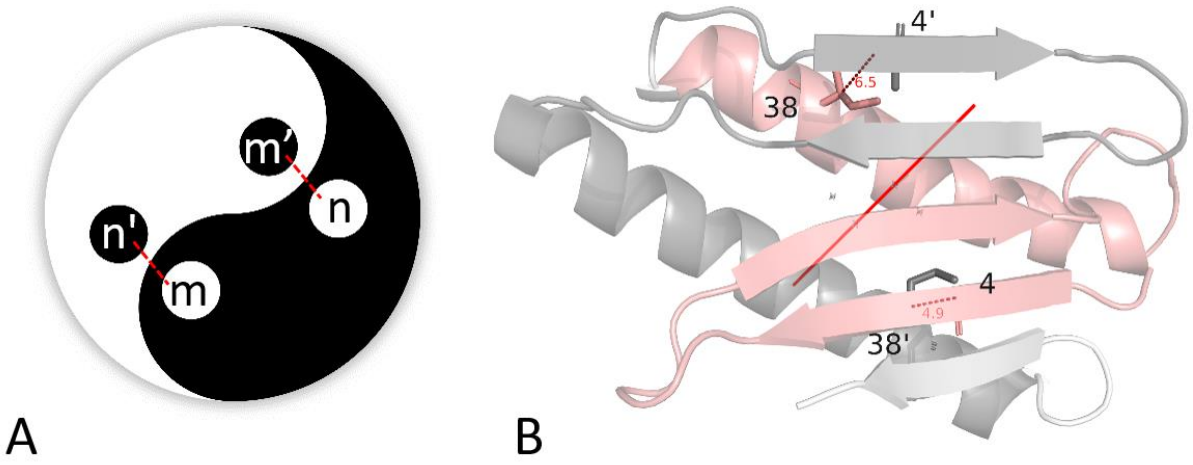

Figure 3. (A) A scheme of contacts in the interacting repetitive units having C2 symmetry. (B) The 3D structure of cysteine protease (Peptidase_Prp) domain with C2 symmetry (PDB entry 2idl:A). Pairs of contacts between residues 4-95 and 58-45, which in the .3D alignment have numbers 4-38' and 4'-38, where 4, 38 are positions of these residues in the first unit and 4', 38' in the second one. If protein had at least one such pair of contacts, it was considered to have C2 symmetry.

The program was written in Python. Input data includes PDB entry with chain ID and numbers of the first and last residues of the repeats. The repetitive structural units are superposed by using TM-align program (Zhang and Skolnick, 2005). The inter-unit contacts were detected by using Contact Map Explorer based on Python library MDTraj (https://contact-map.readthedocs.io/en/latest/index.html). The contact was counted when the distance between two Cβ atoms (or Cα in the case of glycine) was less than 1.0 nm. Our program "C2_contacts" is available at https://github.com/mlkndt/c2

3.4 Prevalence of C2 symmetry in 2RUT structures.

CE-Symm program detects the C2 symmetry in 582 structures of 1198 proteins from the dataset (Supplementary Data 1). Of these 582 structures, our algorithm found the C2 symmetry in 557 structures, that is, in 95.7% of cases. Additionally, the new algorithm detected the C2 symmetry in 249 other structures. Our manual examination of these structures showed that indeed the vast majority of them have C2 symmetry. By summing up all structures having C2 in accordance with either CE-Symm program or our algorithm we got 831 structures, which represents 69,4% (Fig 4). The remaining structures were detected as asymmetrical or having another type of symmetry. The prevalence of the 2RUT structures among all TR structures taken together with the dominance of the C2 symmetry in the 2RUT structures agrees well with the previous observation that the C2 symmetry is the most frequent in TR structures (Myers-Turnbull et al., 2014).

The C2 symmetry prevalence for 2RUT structures (Fig 4) is especially noticeable in the structures with the repeat length less than 55 residues, which usually fold into one structural domain. The preservation of the 2RUT structures can be attributed to the closed C2 symmetry, which imposes structural constraints on the expansion of repeats.

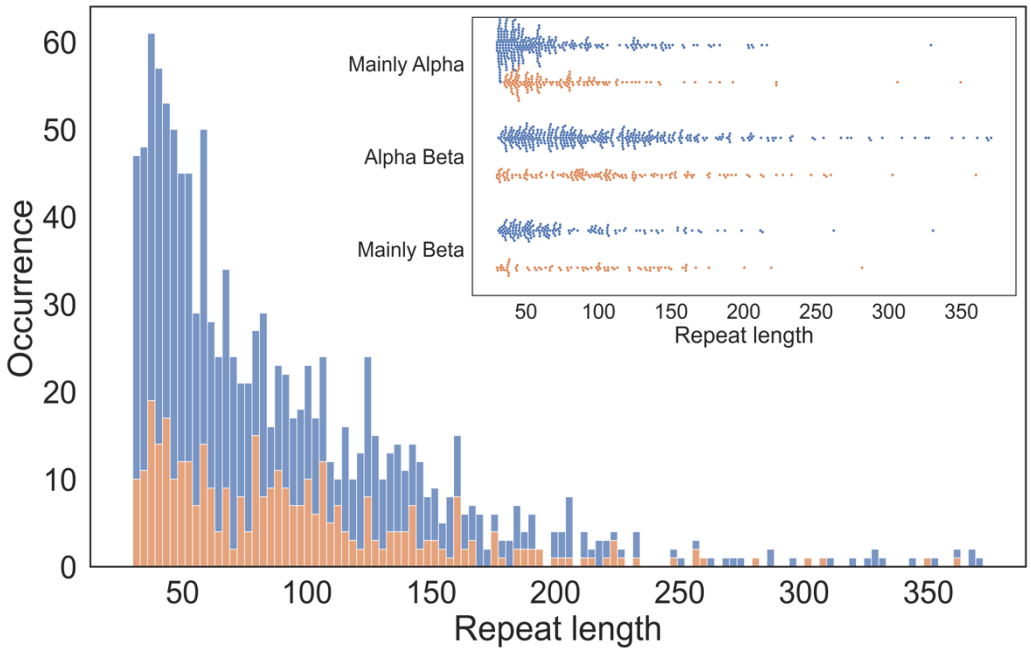

Figure 4. Occurrence of 2RUT structures with C2 in comparison with the other symmetries depending on the repeat length (starting from 30 residues). C2-symmetrical structures are shown in

blue, while others are depicted in orange. An insert with swarm plot shows the distribution of secondary structures depending on the repeat length and symmetry.

3.5 Identification of swapping in 2RUT structures

The 2RUT protein structures with C2 symmetry, like no other structure, creates ideal conditions to exchange parts of the repetitive units with each other, in another word to "swap". Indeed, usually, elements of the secondary structure of a protein domain, being under evolutionary pressure for a long period of time, are well-fitted each other within the 3D structure. When two such units of 2RUT structure interact with each other “face-to-face” in C2 arrangement, a secondary structure element(s) of one unit, which do not fold back to their own domain, can easily find the similar well-fitted nest in the other unit (Fig 5). To a large extent, the tendency of a protein to undergo swapping is governed by thermodynamic factors. In the case of swapped oligomers, the entropic component of the free energy typically tends to disfavor the swapped state in comparison to the monomeric state (Schlunegger et al., 1997). In contrast, the 2RUT structures have a covalent bond linkage between the two repetitive units, resulting in a significant reduction of the entropic effect and promoting swapping. Moreover, this covalent linkage diminishes the influence of protein concentration on swapping, thereby increasing the likelihood that the observed swapping is genuine and not a structural artifact caused by crystallization. Therefore, it was interesting to evaluate the frequency of swapping in these structures.

To identify the "swapping" we developed an algorithm, which uses the residue-residue contact maps (Fig 5). The algorithm was slightly different for the structures with repeat length less and more than 55 residues. In the latter case, of two structural domains, the residues of the swapping part do not have contacts with the residues of their own repeat, but they have contacts with the residues of another repeat. As an example, we show in Figure 5 the residue-residue contact map of the L-arabinose-binding protein structure (pdb code 1abe:A). Residues of the last α-helix (positions 104-124) of the first repeat contact only residues of the second repeat, and vice versa. This situation is clearly visible if to follow a vertical lane at positions 104-124 on the contact map (Fig 5). The lower part of the lane does not have any dots (contacts) while the upper part of the second repeat is full of dots. One can see the opposite pattern in the lane 251-271 corresponding to the equivalent α-helix of the second repeat. Thus, our algorithm detected residue contacts, differentiating between those made within the same repeat and those involving residues from another repeat. A 2RUT structure considered to have swapping if one of its repeats contains more than 15 residues that exclusively form contacts with the other repeat.

When the repeat length is less than 55 residues, two repetitive units form a single structural domain and the residues have contacts with both their own and neighboring repeats. As illustrated on the contact map of one such domain formed by two repetitive units (pdb code 1j27:A), the inter-repeat contacts are concentrated in the top left and bottom right areas. The algorithm detects well swapping when more than 10 residues from one repeat have contacts with the other repeat.

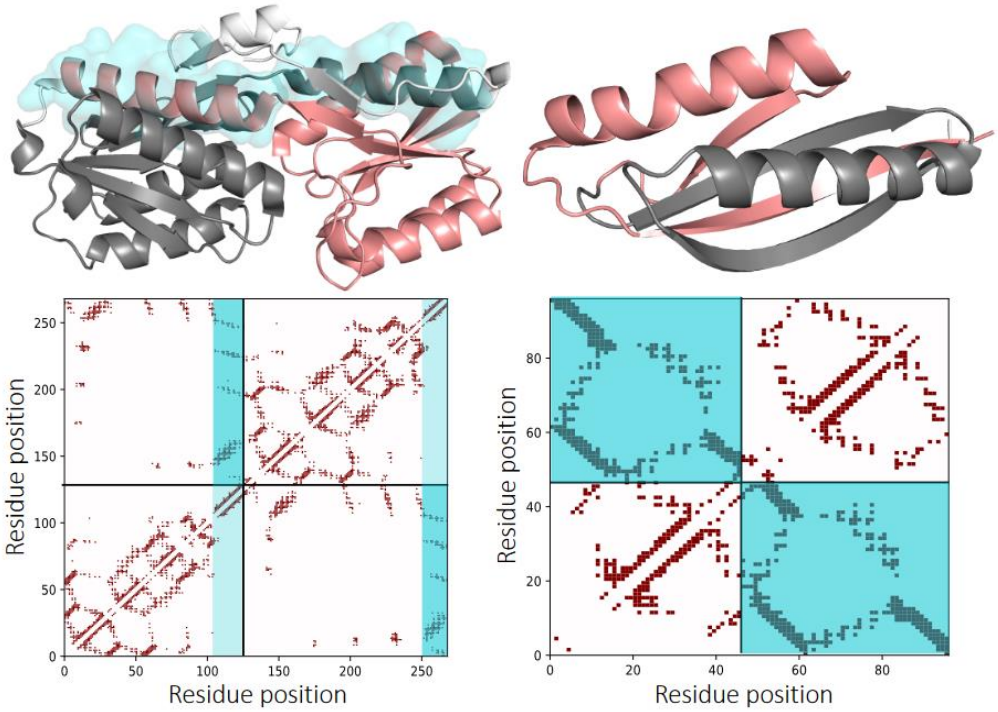

Figure 5. Examples of the structures and their contact maps for swapping identification. Each map is divided into four squares containing intra- and inter-repeat contacts of two repetitive units. Regions with contacts of interest are shown in blue: on the left, these are regions where one repeat has contacts only with another one for the structures with the repeat length over 55 residues (PDB 1abe:A) and, on the right, these are regions where one repeat has numerous contacts with another one for structures with the repeat length less than 55 residues (PDB 1j27:A). The source code of C2_swapping program is available at https://github.com/mlkndt/c2.

Our analysis revealed that 2RUT structures with C2 symmetry have a very high percentage of the swapping. Among 832 structures with C2 symmetry, 269 have swapping that represent 32%. This tendency is especially noticeable for two-fold structures with repeats less than 55 residues where swapping is presented in 36% of structures. For comparison, 367 2RUT structures, which do not have C2 symmetry have only 9% of cases with of the swapping.

3.6 Clustering of 2RUT structures

Although our dataset was filtered to remove structures with high sequence identity, a significant number of similar structures with the same CATH or Pfam IDs remain. Therefore, for a more synthetic representation of the structures we performed their clustering by using variables from Supplementary Data 1: mean length of repeats, type of internal symmetry, secondary structure, fraction of protein occupied by repetitions, existence of swapping, CAT IDs (topology level).

For the clustering, we used the k-means method~~,~~—which was implemented in klearn.cluster.Kmeans module from the scikit-learn Python library (Pedregosa et al., 2011). The clustering was done separately for two groups of proteins with the repeat length less and more than 55 residues. As a result, all structures were grouped into 100 clusters, 40 and 60 of which represent repeats less and more than 55 residues, respectively. The clusters are shown in Supplementary Data 2 (Fig S2). The code for its implementation, which also includes pre-processing and data normalization is presented at https://github.com/mlkndt/c2.

We also performed clustering of our dataset by Foldseek software, which enables fast and sensitive comparisons of large sets of protein structures (Kempen et al., 2022). The clustering was made with a threshold of 80% sequence coverage by the alignment and generated 252 clusters with similar structures (Supplementary Data 1). The results of this clustering can be used in a more detailed structural classification of proteins with TRs.

3.7 Classification of 2RUT structures

Although 2RUT structures are the most common among the structures with TRs, they are not included in the current classification (Kajava, 2012; Paladin et al., 2021). The classification is based on the length of the repeats, which is related to the ability of a single repetitive unit to fold into a stable structure. The shorter the repeat length the more of the repeats in the tandem and more chains associated together are needed to stabilize the structure. For example, TRs with the length between 1 and 30 residues require one another to maintain the structure in the form of crystallite (Class I), oligomers (Class II), or elongated structures with the inter-repeat stabilizing interactions (Class III). Although structures from Class IV consist of relatively long repeats (30-60 residues), individually, these repeats are at the borderline in terms of the structure stability and become stable in the context of closed (ring-like) structures. Finally, proteins with the size of repeats over 50 residues (Class V) are large enough to fold independently into stable domains.

The question arises: where is the place of the 2RUT structures in this classification? The 2RUT structures have a typical repeat length between 30 and 200 residues.

Our analysis of the know 2RUT structures revealed that when their repeat size is between 30 and 55 residues they frequently form one common structural domain with the closed C2 symmetry (Fig. 6). By the repeat length and the necessity of stabilizing interactions to fold into the single structured domains, these 2RUT proteins resemble Class IV structures. The other structures are composed of repeats over 55 residues, which, similarly to the repeats of Class V proteins, are able to form structural domains independently. Thus, we conclude that the structures with units between 30 and 55 residues belong to Class IV structures, while the ones with units of more than 55 residues are a part of Class V structures.

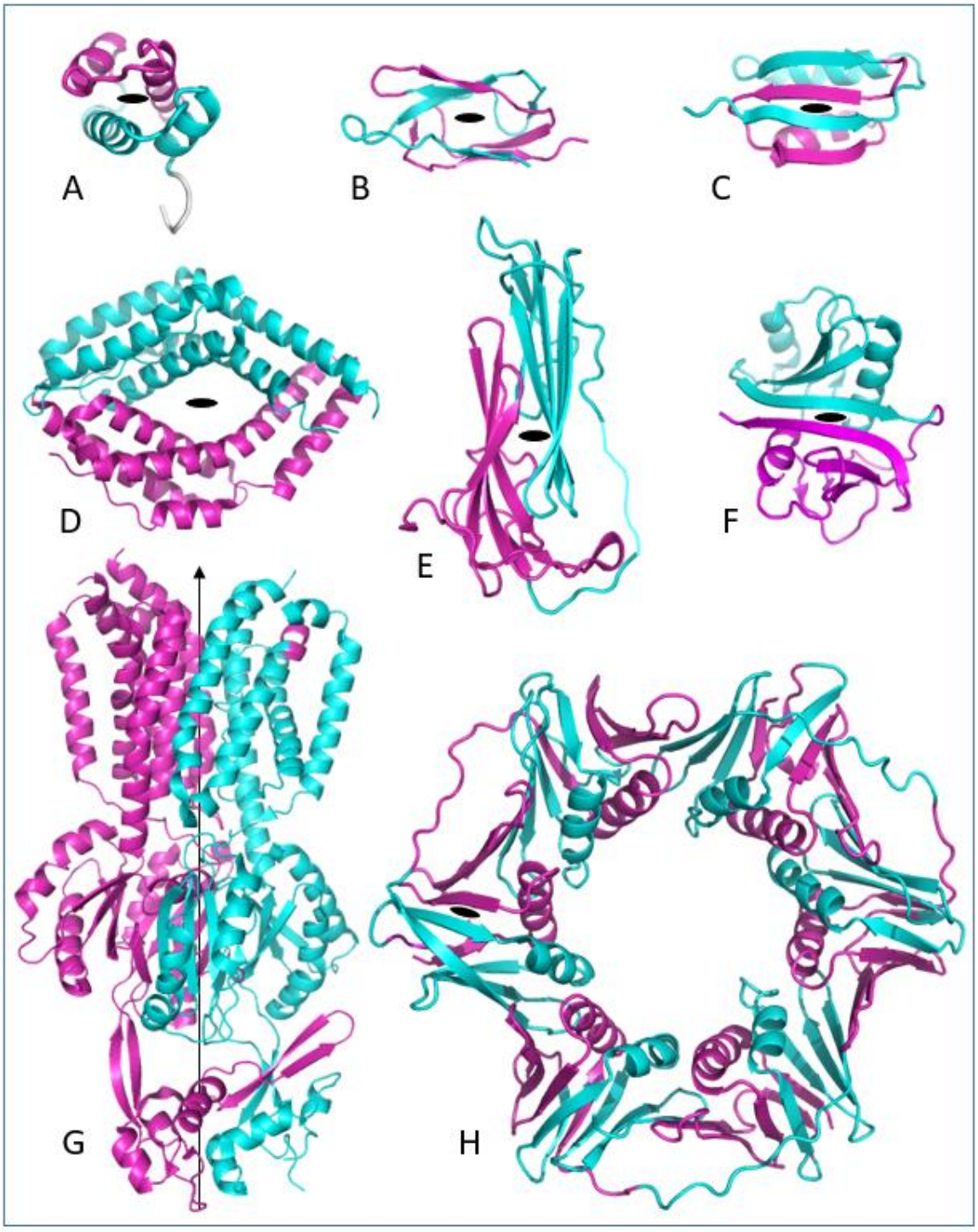

Figure 6. Representative 2RUT structures having C2 symmetry. Position of the twofold axis is represented by a black lens-shaped symbol. Structures with the repeat length less than 55 residues: (A) mainly α-helical, [2Fe-2S] binding domain (PDB entry 3hrd:D), (B) mainly β-structural, biotin attachment domain (PDB entry 1ghj:A), (C) α/β, heavy-metal-associated domain (PDB entry 1cc7:A). Structures with the repeat length over 55 residues: (D) mainly α-helical, conserved histidine α-helical domain (PDB entry 6qv5:A), (E) mainly β-structural, LU domain (Ly-6 antigen/uPAR) of three-finger protein superfamily (PDB entry 6iom:A), (F) α/β, Imp4-like protein (PDB entry 1w94:A), (G) a structure with very long repeat of 480 residues, which contains several structural domains in each repeat, acriflavine resistance protein (PDB entry 4k0e:A), (H) Box repeat structure of heterotrimeric PCNA sliding clamp (PDB entry 2hik:A). In this ring-like structure, three molecules with C2 symmetric beads

interact with one another. The position of one of the twofold axes within the bead is depicted by a black, lens-shaped symbol on the left side of the ring.

3.8 Functions of 2RUT structures

The 2RUT structures also have their own special functions. Unlike many TR proteins, with extended surfaces well suited for protein-to-protein interactions (Andrade et al., 2001, (Peterson et al., 1997; Bork et al., 1994), 2RUT structures usually do not have such a function. In accordance with Pfam annotation, they are often associated with binding to small molecules. Typical examples are ferredoxin, which contains an iron-sulfur cluster and carries out electron transport (Otaka and Ooi, 1987), and heavy metal-associated domains, containing two conserved cysteines involved in binding (Gitschier et al., 1998). Besides metals and their transport, structures with two domains can participate in the binding of other small molecules (Anantharaman and Aravind, 2004), including fatty acid molecules (Schulze-Gahmen et al., 2003).

Among the 2RUT structures, there are many enzymes whose active site is located between two repeating domains (Akey et al., 2010; Chisuga et al., 2017; Haynes et al., 2002; Hofmann et al., 2000; W. Huang et al., 1998; Iyer et al., 2011; Lloyd et al., 2004). For example, the Allantoicase domain, involved in purine cleavage, contains two jelly-roll motifs, each of which has conserved pockets constituting a common active site (Leulliot et al., 2004). On the other hand, repeats do not always have the same effect on enzyme activity. The M16 family peptidase consists of two structurally related domains, but only one of them is an active peptidase, while the other remains inactive. However, the C2 symmetrical structure may help the enzyme to "clamp" the substrate (Taylor et al., 2001). In the case of phosphofructokinases, involved in the regulation of the glycolytic pathway, one domain of the structure, in addition to the binding site, contains an allosteric site that affects the activity of the enzyme (Shirakihara and Evans, 1988).

Occasionally, 2RUT structures exhibit functions that are closely tied to their shapes. For example, the Tli4 family domain, which is part of the effector-immunity Tle-Tli pair in some bacteria, has two domains that together have shape similar to a crab claw. It is used to grasp the Tle domain to prevent its activation (Lu et al., 2014). In the proteins of the TBP family that are required to initiate the transcription process by RNA polymerases the two units form a saddle-like shape, which interacts with transcription initiation factors and regulatory proteins after TATA-box binding (Nikolov et al., 1992).

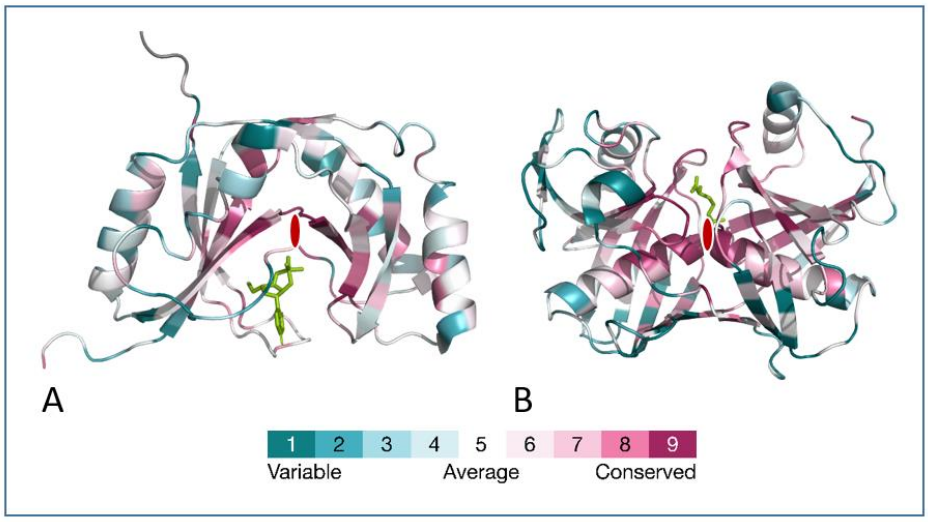

Figure 7. Examples of C2 binding sites. (A) Cyclic phosphodiesterase-like protein domain (PDB entry 1jh7:A), where the central water-filled cavity forms the enzyme's active site (Hofmann et al., 2000). (B) Diaminopimelate epimerase (PDB entry 2gke:A), active-site cleft is formed from both domains (Lloyd et al., 2004). ConSurf web server (Ashkenazy et al., 2016) was used for estimation of the evolutionary conservation of residues.

In some cases, C2 structures with two domains can also be incorporated into multi-repeat structures with a "beads-on-a-string" architecture, where the C2 structures represents the beads. The ample binding surfaces of these proteins make them particularly adept at engaging with large macromolecules (Whitley et al., 2013, Arrías et al., 2023).

4. Conclusions

To investigate still overlooked protein structures with two repetitive units, we built a non-redundant dataset of the known 3D structures. It allowed us to analyze these structures for the occurrence depending on repeat length, internal symmetry, swapping, annotation in Pfam and CATH databases, and functional roles. The analysis revealed that such structures are the most widespread among the TR proteins emphasizing their particular importance.

Depending on the repeat length, all 2RUT structures can be subdivided into two groups: (i), units of less than about 55 residues assemble into one structural domain (Fig. 6A-C) and (ii) the repeats over approximately 55 residues fold in two structural domains (Fig. 6 D-F). The repeats of over 200 residues are rare but if one does occur, it often contains multiple structural domains within that repeat (Fig 6G).

The vast majority of these structures exhibit two-fold symmetry (C2) and this finding provides insight into why the first repeat, which could theoretically continue to multiply at the DNA level, becomes fixed in the process of evolution. Indeed, the closed C2 arrangement provides a structural restriction for the subsequent expansion of the repeats. Due to the C2 symmetry arrangement, two domains cannot provide the necessary binding interface for a third domain, rendering it superfluous. Thus, one can suggest an evolutionary scenario where after the emergence of the second structural domain which is able to interact with the first one in the C2 structure, the third such domain will no longer be able to join the first two domains in a similar manner. Further multiplication is possible if the C2 structure with two repetitive units duplicates as a whole generating tandems with an even number of repeats. Several such structures were observed, including Box-repeat structures (Fig 6H) (Arrías et al., 2023).

The other specific characteristic of the 2RUT structures with C2 symmetry is an extremely high percentage of swapping. Swapping enhances the fidelity of the face-to-face interaction of the two repetitive units through a kind of “hugging”.

The fact that a given symmetrical structure was fixed in evolution does not imply the structural but rather functional reasons when the duplication of a region within a gene imparts a new function, which is beneficial for the organism fitness. Indeed, many 2RUT proteins have very specific functions in comparison to the other TR-containing proteins. The duplicated domains with the C2 symmetry provide structural basis for the emergence of new binding sites for small molecules at and around the axe of two-fold symmetry (Fig. 7). Therefore, these proteins are often associated with binding to small molecules including metals, and their subsequent transport or chemical transformations.

In some other cases, one may assume an evolutionary scenario when duplication of an ancient gene leads to a protein with two domains retaining their original function, and then, this protein undertakes a kind of subfunctionalization with one of the domains losing enzymatic activity for a new function. For example, in phosphofructokinases (PDB 3opy:B) (Shirakihara and Evans, 1988) only one domain has an enzyme activity and another one is responsible for allosteric effect of the enzyme.

It is worth mentioning that sometimes the C2 structures with two domains can also be a part of multi-repeat structures with a "beads on a string" architecture where the two domain structures represent the beads. These proteins have extended binding surfaces suitable for interactions with large macromolecules. For example, in Box-repeat proteins, two-three molecules with such C2 symmetrical beads interact with each other to form ring-like bracelets (see Arrías et al., 2023).

Finally, our analysis led to the classification of these structures within the general structural classification of TR proteins. We suggest that the 2RUT structures with repeats between 30 and 55 residues belong to Class IV structures, while the ones with units over 55 residues are a part of Class V structures. Given that this type of structures is the most abundant among TR proteins, inclusion of them in the RepeatsDB (Paladin et al., 2021) would enhance the database's comprehensiveness.

Acknowledgments

This work was supported by REFRACT MSCA RISE project within European Union's Horizon 2020 research and innovation programme under grant agreement No 823886, as well as from ML4NGP (CA21160), supported by COST (European Cooperation in Science and Technology) under the EU Framework Programme Horizon Europe.

References

Akey, D. L., Razelun, J. R., Tehranisa, J., Sherman, D. H., Gerwick, W. H., & Smith, J. L. (2010). Crystal Structures of Dehydratase Domains from the Curacin Polyketide Biosynthetic Pathway. *Structure*. https://doi.org/10.1016/j.str.2009.10.018

Anantharaman, V., & Aravind, L. (2004). The SHS2 module is a common structural theme in functionally diverse protein groups, like Rpb7p, FtsA, GyrI, and MTH1598/Tm1083 superfamilies. *Proteins: Structure, Function and Genetics*. https://doi.org/10.1002/prot.20140

Andrade, M. A., Perez-Iratxeta, C., & Ponting , C. P. (2001). Protein repeats: structures, functions, and evolution. *Journal of Structural Biology*. https://doi.org/10.1006/jsbi.2001.4392

Arrías P.N., Monzon A.M., Clementel D., Mozaffari S., Piovesan D., Kajava A.V., Tosatto S.C.E. (2023) The repetitive structure of DNA clamps: An overlooked protein tandem repeat. *Journal of Structural Biology*. https://doi.org/10.1016/j.jsb.2023.108001

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research*. https://doi.org/10.1093/NAR/GKW408

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. In *Nucleic Acids Research*. https://doi.org/10.1093/nar/28.1.235

Bliven, S. E., Lafita, A., Rose, P. W., Capitani, G., Prlić, A., & Bourne, P. E. (2019). Analyzing the symmetrical arrangement of structural repeats in proteins with CE-Symm. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1006842

Bork, P., Holm, L., & Sander, C. (1994). The immunoglobulin fold: Structural classification, sequence patterns and common core. In *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.1994.1582

Buard, J., & Vergnaud, G. (1994). Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *EMBO Journal*. https://doi.org/10.1002/j.1460-2075.1994.tb06619.x

Chisuga, T., Miyanaga, A., Kudo, F., & Eguchi, T. (2017). Structural analysis of the dual-function thioesterase SAV606 unravels the mechanism of Michael addition of glycine to an α,β-unsaturated thioester. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.M117.792549

Do Viet, P., Roche, D. B., & Kajava, A. V. (2015). TAPO: A combined method for the identification of tandem repeats in protein structures. In *FEBS Letters*. https://doi.org/10.1016/j.febslet.2015.08.025

Gitschier, J., Moffat, B., Reilly, D., Wood, W. I., & Fairbrother, W. J. (1998). Solution structure of the fourth metal-binding domain from the Menkes copper-transporting ATPase. *Nature Structural Biology*. https://doi.org/10.1038/nsb0198-47

Goodsell, D. S., & Olson, A. J. (2000). Structural symmetry and protein function. In *Annual Review of Biophysics and Biomolecular Structure*. https://doi.org/10.1146/annurev.biophys.29.1.105

Haynes, C. A., Koder, R. L., Miller, A. F., & Rodgers, D. W. (2002). Structures of nitroreductase in three states. Effects of inhibitor binding and reduction. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.M111334200

Heringa, J. (1998). Detection of internal repeats: How common are they? *Current Opinion in Structural Biology*. https://doi.org/10.1016/S0959-440X(98)80068-7

Hofmann, A., Zdanov, A., Genschik, P., Ruvinov, S., Filipowicz, W., & Wlodawer, A. (2000). Structure and mechanism of activity of the cyclic phosphodiesterase of Appr>p, a product of the tRNA splicing reaction. *EMBO Journal*, *19*(22), 6207–6217. https://doi.org/10.1093/emboj/19.22.6207

Huang, W., Jia, J., Edwards, P., Dehesh, K., Schneider, G., & Lindqvist, Y. (1998). Crystal structure of β-ketoacyl-acyl carrier protein synthase II from E.coli reveals the molecular architecture of condensing enzymes. *EMBO Journal*. https://doi.org/10.1093/emboj/17.5.1183

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq003

Iyer, L. M., Zhang, D., Rogozin, I. B., & Aravind, L. (2011). Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkr691

Jorda, J., Xue, B., Uversky, V. N., & Kajava, A. V. (2010). Protein tandem repeats - the more perfect, the less structured. *FEBS Journal*. https://doi.org/10.1111/j.1742-464x.2010.07684.x

Kajava, A. V. (2001). Review: Proteins with repeated sequence - Structural prediction and modeling. In *Journal of Structural Biology*. https://doi.org/10.1006/jsbi.2000.4328

Kajava, A. V. (2012). Tandem repeats in proteins: From sequence to structure. *Journal of Structural Biology*. https://doi.org/10.1016/j.jsb.2011.08.009

Kempen, M. van, Kim, S. S., Tumescheit, C., Mirdita, M., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2022). Foldseek: fast and accurate protein structure search. *BioRxiv*.

Leulliot, N., Quevillon-Cheruel, S., Sorel, I., Graille, M., Meyer, P., Liger, D., Blondeau, K., Janin, J., & Van Tilbeurgh, H. (2004). Crystal structure of yeast allantoicase reveals a repeated jelly roll motif. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.M401336200

Lloyd, A. J., Huyton, T., Turkenburg, J., & Roper, D. I. (2004). Refinement of Haemophilus influenzae diaminopimelic acid epimerase (DapF) at 1.75 Å resolution suggests a

mechanism for stereocontrol during catalysis. *Acta Crystallographica Section D: Biological Crystallography*. https://doi.org/10.1107/S0907444903027999

Lu, D., Zheng, Y., Liao, N., Wei, L., Xu, B., Liu, X., & Liu, J. (2014). The structural basis of the Tle4-Tli4 complex reveals the self-protection mechanism of H2-T6SS in Pseudomonas aeruginosa. *Acta Crystallographica Section D: Biological Crystallography*. https://doi.org/10.1107/S1399004714023967

Marcotte, E. M., Pellegrini, M., Yeates, T. O., & Eisenberg, D. (1999). A census of protein repeats. *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.1999.3136

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa913

Myers-Turnbull, D., Bliven, S. E., Rose, P. W., Aziz, Z. K., Youkharibache, P., Bourne, P. E., & Prlić, A. (2014). Systematic detection of internal symmetry in proteins using CE-symm. *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2014.03.010

Nikolov, D. B., Hu, S. H., Lin, J., Gasch, A., Hoffmann, A., Horikoshi, M., Chua, N. H., Roeder, R. G., & Burley, S. K. (1992). Crystal structure of TFIID TATA-box binding protein. *Nature*. https://doi.org/10.1038/360040a0

Otaka, E., & Ooi, T. (1987). Examination of protein sequence homologies: IV. Twenty-seven bacterial ferredoxins. *Journal of Molecular Evolution*. https://doi.org/10.1007/BF02099857

Paladin, L., Bevilacqua, M., Errigo, S., Piovesan, D., Mičetić, I., Necci, M., Monzon, A. M., Fabre, M. L., Lopez, J. L., Nilsson, J. F., Rios, J., Menna, P. L., Cabrera, M., Buitron, M. G., Kulik, M. G., Fernandez-Alberti, S., Fornasari, M. S., Parisi, G., Lagares, A., … Tosatto, S. C. E. (2021). RepeatsDB in 2021: Improved data and extended classification for protein tandem repeat structures. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa1097

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

Peterson, A. J., Kyba, M., Bornemann, D., Morgan, K., Brock, H. W., & Simon, J. (1997). A domain shared by the Polycomb group proteins Scm and ph mediates heterotypic and homotypic interactions. *Molecular and Cellular Biology*. https://doi.org/10.1128/mcb.17.11.6683

Richard, F. D., & Kajava, A. V. (2014). TRDistiller: A rapid filter for enrichment of sequence datasets with proteins containing tandem repeats. *Journal of Structural Biology*. https://doi.org/10.1016/j.jsb.2014.03.013

Saupe, S., Turcq, B., & Bégueret, J. (1995). A gene responsible for vegetative incompatibility in the fungus Podospora anserina encodes a protein with a GTP-binding motif and Gβ homologous domain. *Gene*. https://doi.org/10.1016/0378-1119(95)00272-8

Schlunegger, M.P., Bennett, M.J., & Eisenberg, D. (1997) Oligomer formation by 3D domain swapping: A model for protein assembly and misassembly. *Advances in Protein Chemistry*. 50, 61-122, https://doi.org/10.1016/S0065-3233(08)60319-8

Schulze-Gahmen, U., Pelaschier, J., Yokota, H., Kim, R., & Kim, S. H. (2003). Crystal structure of a hypothetical protein, TM841 of Thermotoga maritima, reveals its function as a fatty acid-binding protein. *Proteins: Structure, Function and Genetics*. https://doi.org/10.1002/prot.10305

Shirakihara, Y., & Evans, P. R. (1988). Crystal structure of the complex of phosphofructokinase from Escherichia coli with its reaction products. *Journal of Molecular Biology*, *204*(4), 973–994. https://doi.org/10.1016/0022-2836(88)90056-3

Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S. M., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., Le Cornu, S., Lam, S. D., Berka, K., Varekova, I. H., Svobodova, R., Lees, J., & Orengo, C. A. (2021). CATH: Increased structural coverage of functional space. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa1079

Simon, M., & Hancock, J. M. (2009). Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biology*. https://doi.org/10.1186/gb-2009-10-6-r59

Taylor, A. B., Smith, B. S., Kitada, S., Kojima, K., Miyaura, H., Otwinowski, Z., Ito, A., & Deisenhofer, J. (2001). Crystal structures of mitochondrial processing peptidase reveal the mode for specific cleavage of import signal sequences. *Structure*. https://doi.org/10.1016/S0969-2126(01)00621-9

Tompa, P. (2003). Intrinsically unstructured proteins evolve by repeat expansion. In *BioEssays*. https://doi.org/10.1002/bies.10324

Whitley, M. J., Furey, W., Kollipara, S., & Gronenborn, A. M. (2013). Burkholderia oklahomensis agglutinin is a canonical two-domain OAA-family lectin: Structures, carbohydrate binding and anti-HIV activity. *FEBS Journal*. https://doi.org/10.1111/febs.12229

Zhang, Y., & Skolnick, J. (2005). TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gki524