The repetitive structure of DNA clamps: An overlooked protein tandem repeat

Paula Nazarena Arrías, Alexander Miguel Monzon, Damiano Clementel, Soroush Mozaffari, Damiano Piovesan, Andrey V Kajava, Silvio C.E. Tosatto

The repetitive structure of DNA clamps: an overlooked protein tandem repeat

Paula Nazarena Arrías[1], Alexander Miguel Monzon[2], Damiano Clementel[1], Soroush Mozaffari[1], Damiano Piovesan[1], Andrey V. Kajava[3], Silvio C.E. Tosatto[1#]

[1] *Department of Biomedical Sciences, University of Padova, via U. Bassi 58/b, 35121 Padova, Italy*
[2] *Department of Information Engineering, University of Padova, via Giovanni Gradenigo 6/B, 35131 Padova, Italy*
[3] *Centre de Recherche en Biologie cellulaire de Montpellier (CRBM), UMR 5237 CNRS, Université Montpellier, 1919 Route de Mende, Cedex 5, 34293 Montpellier, France*

[#] *To whom correspondence should be addressed. E-mail: silvio.tosatto@unipd.it; Tel: +39 049 827 6269*

ABSTRACT

Tandem repeat proteins (TRPs) are a widespread class of non-globular proteins that contain repetitive stretches of amino acids. Structured tandem repeats in proteins (STRiPs) are a specific type of tandem repeats, characterized by a modular three-dimensional structure arrangement. The majority of STRiPs adopt solenoidal structures, but with the increased availability of experimental structural data and top-quality structural predictive models more STRiPs folds can be characterized. Here, we describe an overlooked STRiPs fold present in the DNA sliding clamp processivity factors, which has eluded classification although structural data has been available since the late 1990s. “Box” repeats comprise a fold inside class V of “beads-on-a-string” STRiPs which has been added to the RepeatsDB database, which now provides structural annotation for 66 of these proteins belonging to different organisms, including viruses.

KEYWORDS

Tandem repeat protein, box repeats, structured tandem repeat

INTRODUCTION

Tandem repeat proteins (TRPs) are a ubiquitous type of non-globular proteins characterized by repetitive sequence elements arranged in tandem (Kajava & Tosatto, 2018). The length of these

repetitive stretches, or “repetitive units”, can range from just a few residues to more than a hundred. This distinct organization can generate a modular 3D protein structure made up of repetitions of the same structural unit denominated STRiPs (Structural Tandem Repeats in Proteins). The repetitive units are defined as the smallest structural building block that make up the repetitive region (Di Domenico et al., 2014). Repetitive regions however, are usually not perfect and can include insertions, *i.e.* segments that do not belong to the repetitive units, which can be found inside the units or between them.
TPRs are reported to be highly prevalent in eukaryotes, but they are also present in bacteria and archaea, as well as in some viruses (Marcotte et al., 1999; Delucchi et al., 2020; Moore et al., 2008).

Kajava (Kajava, 2012) proposed a classification for STRiPs based on their architecture and the length of their units, grouping tandem repeat protein structures into five distinct classes. Class I groups proteins that form crystalline aggregates, in which the repetitive regions have very short repeat units (1 or 2 amino acids). Class II gathers fibrous structures that require interchain interactions for stabilization, and in which the repeats have a length of 3 to 7 residues. Class III is mainly composed of elongated structures (mostly different types of solenoids), with repeats in the range from 5-45 amino acids. The units in this class require one another to maintain the structure. Class IV of closed structures, has a similar repeat length to Class III (each repeat averaging 30-60 amino acids in length), but in contrast to elongated repeats, have a fixed number of units due to the circular nature of the structure, although they still need one another to maintain the stability. Lastly, Class V is mainly dominated by structures of different “beads on a string” repeats, in which the units have average lengths of over 50 residues. These longer values in repeat length, makes it possible for each unit to fold into small “globular” domains loosely connected to each other, like the repeats present in, for example, different types of zinc fingers and annexin repeats. However, the increased structural data deposited into the Protein Data Bank has led to the identification of other types or structures that belong to Class V. Included within these structures are spectrin repeats (IPR002017) and sushi repeats (IPR000436), among others.

Identification, classification and annotation of new types of STRiPs is a continuous effort. The RepeatsDB database (Paladin et al., 2021) collects experimental STRiPs information, and provides both classification (based on Kajava’s proposal) and annotation of tandem repeat protein structures.

DNA replication is a crucial process in all domains of life, including some viruses, which is carried out by a multi-protein complex denominated DNA replisome. A quintessential protein of the replisome is the DNA polymerase (DNApol). DNA polymerases synthesize complementary DNA in a 5' to 3' direction, and present outstanding fidelity which assures precise replication of the genomes (Lehman et al., 1958). In contraposition to their high fidelity, DNA polymerases, in general, have very low processivity. Processivity refers to the ability of the DNApol to synthesize DNA continuously on a template without dissociating from it, and it is usually measured in terms of the number of nucleotides that are incorporated into the growing DNA molecule per binding event (Zhuang & Ai, 2010). For the complete replication of large DNA genomes, DNApol processivity is not enough. However, this has been made possible due to the existence of other proteins; the processivity factors (Mace & Alberts, 1984). A specific type of processivity factors, known as DNA sliding clamps, are proteins that enhance the processivity of the DNApol by holding it together with the DNA.

Although structural information of DNA sliding clamps has been around since the early 1990s (Kong et al., 1992), they have not been previously classified as containing STRiPs. Here we set out to provide a classification for these processivity factors following Kajava's scheme, comparing their structures, sequences and distribution across the domains of life.

METHODS

The RepeatsDB database (Paladin et al., 2021) was used to initially retrieve a dataset of protein structures with "Box" repeats. In addition, DNA clamp protein sequences with experimental structures were retrieved from UniProt (The UniProt Consortium, 2023) by using the CATH superfamily identifier 3.70.10.10 which contains the fold denominated "box". These initial sets of structures were grouped by protein sequence with the UniProt accession number, and by protein family/domain with Pfam (Mistry et al., 2021). The identification of repeated units and insertions was manually performed following the same protocol of biocuration in the RepeatsDB database. A total of 66 PDB structures were annotated, representing 64 different protein sequences and Pfam families/domains (Supplementary Table 1).

Structural similarity between repeat units was calculated by performing pairwise structural alignments with the software TM-align (Zhang & Skolnick, 2005). The multiple sequence alignment of the units were derived from the aligned residues on the structural alignment, and

visualized with the software JalView (Waterhouse et al., 2009). The alignment consensus was calculated using the EMBOSS Cons tool (EMBL-EBI). The clustering analysis was performed using Scikit-learn and the algorithm DBSCAN (Pedregosa et al., 2011; Ester et al., 1996).
FoldSeek server (van Kempen et al., 2022) was used to search the PDB for similar folds to Box repeats.
PyMOL Molecular Graphics System was used for protein structure representation (Schrödinger, LLC, 2015).

RESULTS AND DISCUSSION

DNA clamps are composed of structural tandem repeats

Structural architecture and unit definition

Most DNA clamps form ring-shaped oligomers with six domains and have been previously described to have 6-fold pseudosymmetry (Oakley, 2016). The domains are superimposable with each other and they are composed of two □-helices and eight β-strands. However, looking into the arrangement of the secondary structure elements in each domain it is possible to define two repetitions of β□βββ modules. These repetitions, or units, are arranged in tandem (Figure 1A). Although the structural similarity of the units belonging to the same protein is high, sequentially this is not always the case (Figure 1B and 1C).
The overall arrangement of the units gives rise to a structure denominated as “box” in CATH/Gene3D database, and all the structures of DNA clamps analyzed (Table 1) belong to the superfamily 3.70.10.10 (Dawson et al., 2017).

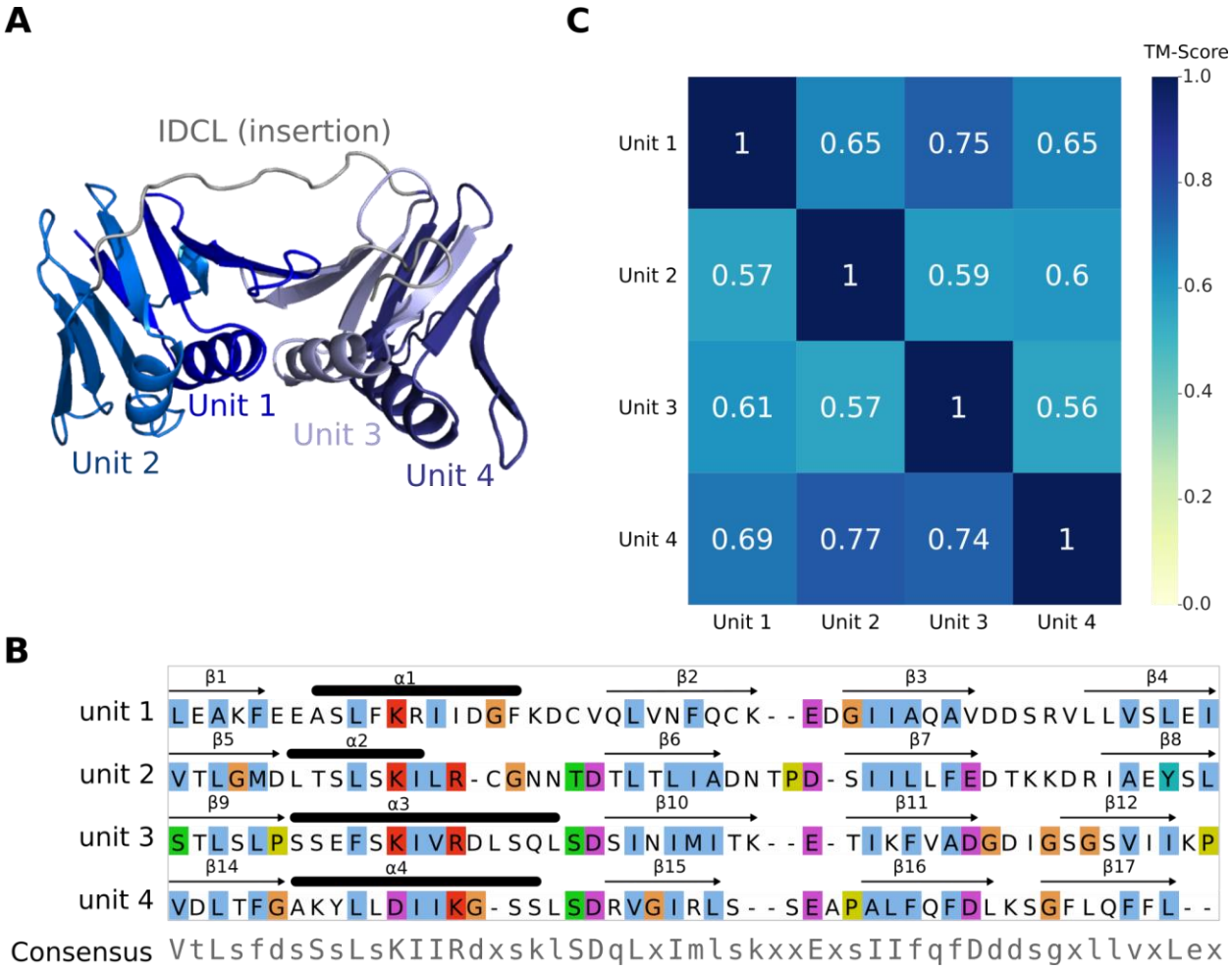

Figure 1: Repetitive units definition and organization in DNA clamps. (A) *Unit organization and insertion in PDB 1plqA.* ***(B)*** *Unit multiple sequence alignment. Arrows represent* β*-strands and black cylinders represent* □*-helices.* ***(C)*** *Unit structural similarity matrix (TM-score) for PDB 1plqA.*

Insertions

Tandem repeats are usually not perfect, and their modular structure makes them prone to deletions, duplications and insertions. Insertions are elements that do not belong to the repeat units, and can appear inside units or between them. They do not contribute to the stability of the repeat, but they usually have a functional role, such as binding (Paladin et al., 2020). Such is the case of DNA clamps, in which every two units there is a non-repetitive segment, composed by a loop (IDCL, inter-domain connecting loop), which connects the last unit of one domain to the first one of the other (Figure 1A). The interaction of DNA clamps with multiple of their partners, such as p21, Fen1 and DNA ligase, among others, are mediated through residues of the IDCL (Pascal et al., 2006; Sakurai et al., 2005; Warbrick et al., 1995).

The structural organization of DNA sliding clamps varies across the domains of life

Bacterial DNA sliding clamps

The first structural information of these processivity factors came out in 1992 with the determination of *Escherichia coli* beta sliding clamp (coded by gene *dnaN*) structure (Kong et al., 1992). Up to this day, the same protein (UniProt ID P0A988) has been crystallized multiple times, and the structures of other 23 bacterial sliding clamps have been solved (Table 1 and Supplementary Table 1).

Bacterial sliding clamp monomers have about 378 amino acids. They present three distinct Pfam domains; PF00712, PF02768, PF02767 (Figure 2A). The structures of these domains are highly similar, showing an average TM-score of 0.75 (Supplementary Figure 1).

Two monomers assemble in a heat-to-tail fashion to form a donut-shaped homodimer that accommodates double stranded DNA (dsDNA) in a central cavity of around 35 Å in diameter (Kong et al., 1992) (Figure 2B). The dimer interacts with DNA polymerase III and is “loaded” onto DNA by the action of the clamp loader complex.

Eukaryotic PCNA and 9-1-1 clamp

Homo sapiens DNA sliding clamp (UniProt ID P12004), commonly known as proliferating cell nuclear antigen (PCNA), was originally discovered as an antigen reacting with antibodies derived from systemic lupus erythematosus patients’ sera (Miyachi et al., 1978). Homologs in other eukaryotic organisms were described afterwards. Although there is virtually no sequence similarity between eukaryotic and bacterial DNA sliding clamps (Acharya et al., 2021), their structures are highly similar. Each PCNA monomer has an average length of 260 residues, and presents two distinct Pfam domains; PF00705 and PF02747. As in the case of their bacterial counterparts, both domains share structural similarity, with an average TM-score of 0.79 (Supplementary Figure 1), but in contraposition to the DNA sliding clamps of bacteria, PCNA monomers assemble into a trimer (Gulbis et al., 1996). The PCNA trimer increases DNApol processivity by directly interacting with it and holding it to the DNA during chromosome replication.

In addition to PCNA, eukaryotic cells also have another structurally similar DNA sliding clamp (Venclovas & Thelen, 2000) that is involved in replication checkpoint control (S-phase progression, G2/M arrest) and DNA repair that has been denominated the 9-1-1 complex (Parrilla-Castellar et al., 2004). A fundamental difference between PCNA and the 9-1-1 complex is that

the latter is a heterotrimer made up from 3 distinct proteins that in humans are called RAD9 (UniProt ID Q99638), RAD1 (UniProt ID O60671) and HUS1 (UniProt ID O60921).

RAD9 of *Homo sapiens* is a protein of 391 amino acids, almost the average length of the bacterial sliding clamps, but it is structurally similar to PCNA, with the difference that Pfam assigns only one domain to the entire protein (PF04139).

RAD1 and HUS1 have lengths of 280 and 282 amino acids respectively, more similar to the average length of the rest of the eukaryotic PCNAs. As with RAD9, Pfam only assigns one domain for each protein, being PF02144 for RAD1 and PF04005 for HUS1. In the yeast *Saccharomyces cerevisiae,* the proteins DDC1 (UniProt ID Q08949), MEC3 (UniProt ID Q02574) and RAD17 (UniProt ID P48581) make up the 9-1-1 clamp. The structures of 20 different proteins have been deposited into the PDB so far (Table 1 and Supplementary Table 1).

Archaeal PCNA

The DNA sliding clamps of Archaea resemble eukaryotic PCNA. Each monomer has an average length of 248 residues and presents the two distinct domains of eukaryotic PCNA; PF00705 and PF02747. In contrast to eukaryotic PCNA, some archaeans present two (*Thermococcus kodakarensis*) and even three (*Sulfolobus solfataricus* and *Sulfurisphaera tokodaii)* PCNA homologs, which assemble as heterotrimers, or even as a heterotetramer which has been proposed can accommodate in its cavity a Holliday junction (Kawai et al., 2011). There are currently 14 different archaeal PCNAs structures in the PDB (Table 1 and Supplementary Table 1).

Viral processivity factors

Some double stranded DNA viruses and bacteriophages also have proteins that act as processivity factors. Currently, the structure of only 6 of these viral proteins (Table 1 and Supplementary Table 1) has been elucidated.

The protein GP45 of two species of bacteriophages; Mosigvirus RB69 (previously known as Escherichia phage RB69) and Tequatrovirus T4 (also known as T4 bacteriophage) have a length of 228 amino acids. Similarly to eukaryotic and archaeal DNA sliding clamps, GP45 forms a homotrimeric ring, and each monomer has two distinct domains; PF02916, PF09116, which are structurally similar although sequentially there is no apparent similarity.

Human alphaherpesvirus 1 (HHV-1) processivity factor, a protein of 488 amino acids denominated UL42 (UniProt ID P10226), differs from the previously mentioned DNA sliding clamps. Structurally, the N-terminal domain of UL42 adopts a similar conformation as archaeal and

eukaryotic clamps but the C-terminal domain is predicted to be disordered by MobiDB (Piovesan et al., 2018). Pfam assigns a sole domain for the N-terminal region, PF02282, which is a member of the CL0060 clan (same as the other sliding clamps). In contrast to the other sliding clamps, UL42 apparently binds DNA as a monomer and not as a ring-shaped oligomer (Randell & Coen, 2004).

Table 1. **Experimental structures analyzed in this work.** *PDB codes and chain IDs grouped by domain of life.*

Domain of Life							
Bacteria			Eukarya		Archaea		Viruses
6degA	3t0pA	6d46A	3g65A	4ztdA	3hi8A	2hiiA	1b77A
4k74A	4tr6A	6ptvA	3a1jC	6qh1A	5a6dA	2hikC	1czdA
6amqC	6dj8A	2avtA	3a1jB	1plqA	6t8hE	1rwzA	2z0lA
6ap4A	5wceA	2awaA	7wp3E	7bupA	6aigA	1iz4A	3hslX
4tr7A	5x06A	6d47A	3k4xA	7ep8A	1ud9A	3aixA	1t6lA
3p16A	4n96A	6manA	5tupA	7o1eA	3lx1A	3lx2A	1dmlA
5ah2A	4rkiA	4trtA	3p91A	2zvwA	2hiiB	3aixB	
6dm6A	4tr8A	1vpkA	4cs5A	7sh2H			
6d46A	6manA	7evpA	2zvvA	7sh2G			
7rzmA	7evpA	7rzmA	4hk1A	7sh2F			

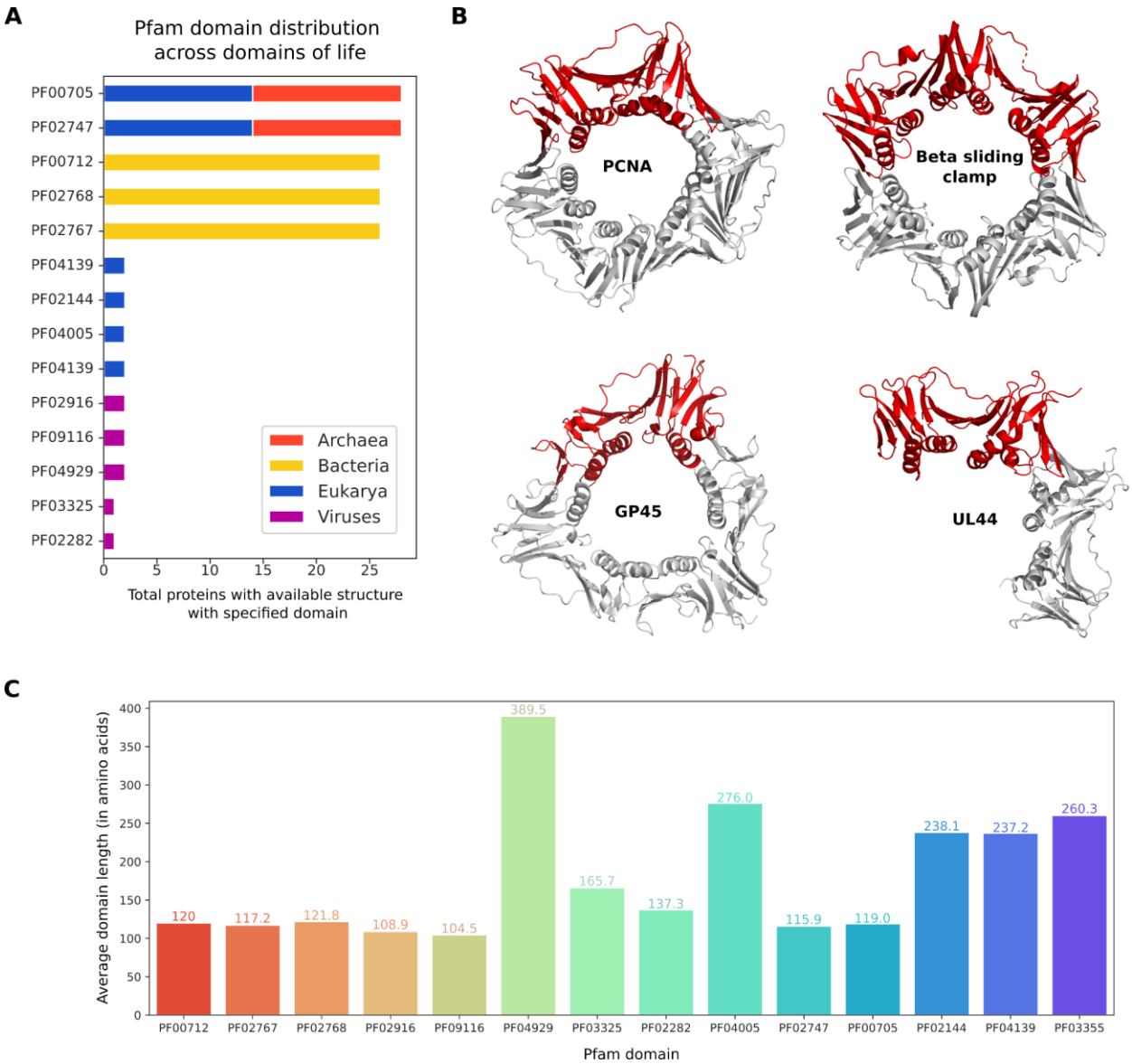

Figure 2: Structural organization of DNA sliding clamps across the domains of life (A) *Pfam domains present in DNA clamps and their distribution across domains of life (including viruses).* ***(B)*** *Oligomerization states of DNA clamps. In red one monomer.* ***(C)*** *Average length of the Pfam domains present in DNA clamps.*

Similarly to HHV-1 UL42, the processivity factor of another herpesvirus, UL44 (UniProt ID P16790) of human betaherpesvirus 5 (HHV-5), a protein of 433 residues, presents a N-terminal domain (PF03325) that is structurally similar to PCNA. However, the crystal structure suggests that it exerts its function as a C-shaped dimer, in which both monomers interact in a head-to-head fashion, instead of a ring formed by the heat-to-tail interaction of the monomers (Appleton et al., 2004). This situation also applies for HHV-8 protein PF-8 (UniProt ID Q77ZG5) (Baltz et al., 2009)

and BMRF1 (UniProt ID P03191) of HHV-4 (Murayama et al., 2009), which have a length of 396 and 404 residues, respectively. The N-terminal domain of both proteins is classified as PF04929.

The length of the monomers defines the number of units within the repetitive region; in the case of eukaryotic, archaeal and viral DNA clamps, the number of units in each monomer is four, while the bacterial clamps have six (Figure 3A and 3C). The average length of each unit is quite homogeneous in the case of Archaea, Bacteria and Eukarya but for viruses this number can vary quite a lot (Figure 3D). However, the number of experimental structures for viral processivity factors available on the PDB is small (Table 1 and Supplementary Table 1).

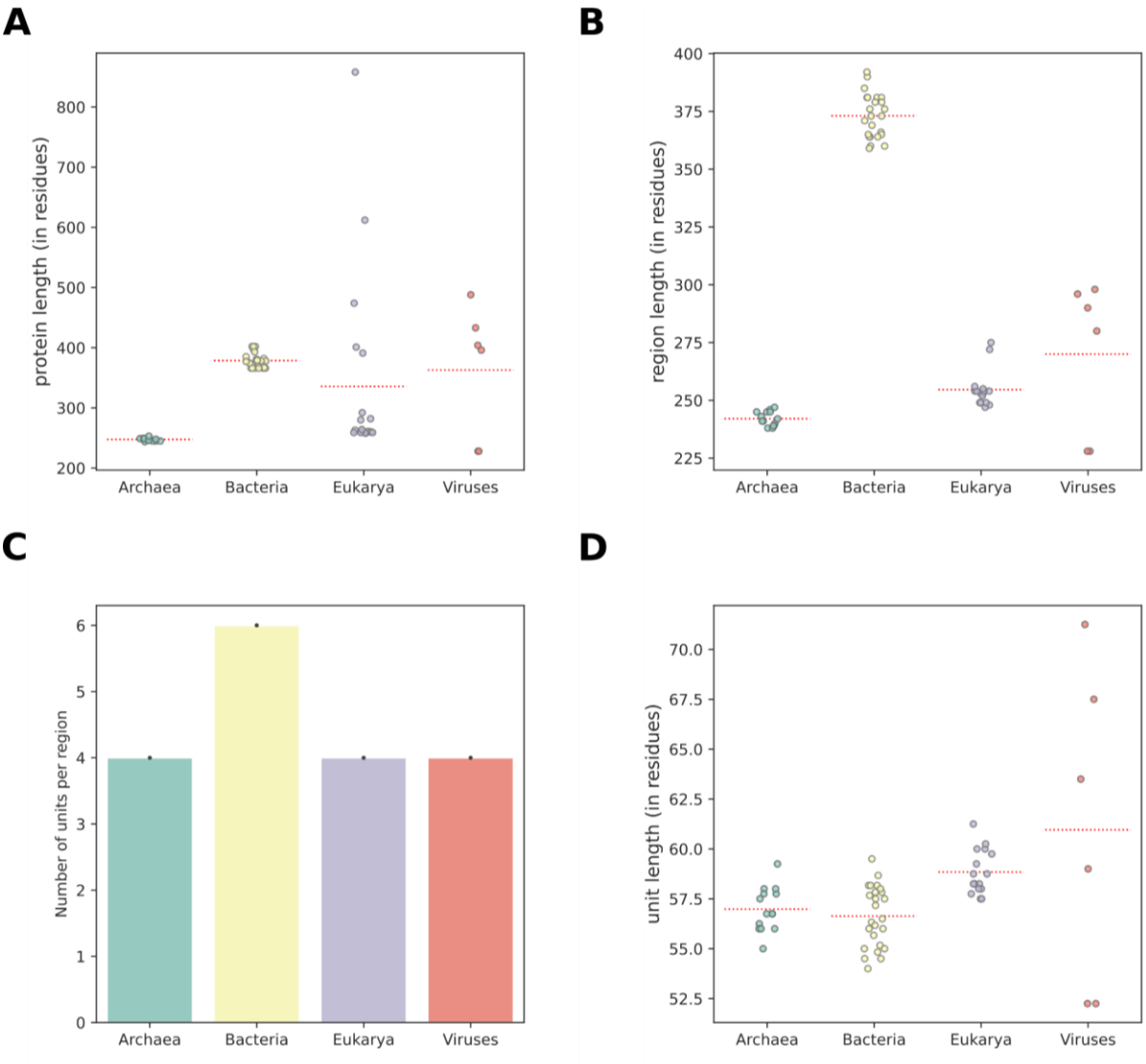

Figure 3: The repeats present in DNA clamps vary across domains of life. (A) *Protein average length in amino acids* ***(B)*** *Repetitive region average length in amino acids* ***(C)*** *Total number of units per region in*

the different domains of life. ***(D)*** *Unit length distribution (in residues). In all cases, the red dotted line represents mean values.*

Structural Tandem Repeat classification of DNA clamps: "Box" repeats

General classification

Box repeats are about 60 residue long. Given the average unit length of the repeats as well as the tight connection between units, box repeats can be classified as belonging to either Class IV or Class V structures. Kajava's classification is based not only on the length of the repeat, but also on the context under which a repetitive unit of such a length can fold into a stable structure. The individual repeats of the Class IV structures cannot form a stable structure and become stable in the context of closed (ring-like) structures. At the same time, the repeats from Class V are large enough to fold independently into stable domains. Thus, evidence about the ability of a single Box repeat to fold or not to fold into the stable structure becomes essential for the classification. We searched the PDB database to find single structural domains similar to the β□βββ modules of Box repeats. Some proteins, such as the Ragulator complex protein LAMTOR4 (Q0VGL1, PDB ID 5VOK_A) and the polyprotein of Hepacivirus C (H9XGD6, PDB ID 4UOI_A) have single domains with the structures (Rasheed et al., 2019; El Omari et al., 2014) similar to one Box repeat, suggesting that Box repeats have a potential to fold independently. Therefore, Box repeat proteins can be assigned to Class V rather than to Class IV.

RepeatsDB classification

RepeatsDB (Paladin et al., 2021) is a database that provides annotation and classification of STRiPs. The classification is based on Kajava's scheme, with the first level of classification being "Class". However, it incorporates 4 extra levels of classification; "Topology" (second level of classification), which is distinguished by the general path of the polypeptide chain and the type of secondary structure in the repeat units, "Fold" (third level of classification), which comprises variants of a certain topology that present structural differences in the number of secondary structure elements, additional structural elements and overall structural arrangement, "Clan" (fourth level of classification) which is a subfold, and lastly "Family" (fifth level of classification), which groups proteins with a common ancestor based on sequence similarity.

Within this classification, Box repeats comprise a unique fold within the Alpha/beta beads topology of Class V. Structural clustering analysis of the regions employing DBSCAN shows that the box repeats from bacteriophages form a separate cluster in the range of 73%-76% of structural similarity and hence we propose they form a separate clan. The level of family has not yet been implemented into the current version of the database, and thus remains unassigned for box repeats.

Eukaryotic PCNA exon arrangement does not exhibit the periodicity of their repeat units

It has been suggested that in eukaryotes, repeat segments could correspond to exons, thus allowing their easy duplication and shuffling (Schaper & Anisimova, 2015). The instances in which the structural symmetry of the tandem repeat regions is also seen in their exon arrangement supports the notion that STRiPs can evolve through duplication of their coding exons. Such is the case of some repeat types, such as ankyrins and leucine-rich repeats (Paladin et al., 2020).
We examined this periodicity in the exon arrangement for the eukaryotic DNA clamps with available structure, and employed the pipeline developed by Paladin and collaborators (Paladin et al., 2020) to produce their “repeat/exon plots”, which allow the visualization of the alignment between the structural units and exons.
We found that for eukaryotic clamps there is not a uniform pattern. For example, both *Saccharomyces cerevisiae* PCNA (P15873) and the three proteins that compose its 9-1-1 complex are encoded by a single exon, while *Homo sapiens* PCNA (P12004) is encoded by 6, but shows a complex pattern (data not shown). *Arabidopsis thaliana* PCNA1 (Q9M7Q7) and PCNA2 (Q9ZW35) are, however, encoded by 4 exons, but these do not correspond to the structural units, just partial parts of them.

CONCLUSION

Here we have provided a classification for the structural tandem repeats present in the DNA clamp processivity factors within Kajava’s scheme. We have manually annotated 66 protein structures from different organisms, including DNA viruses. This new data and structural description will help to increase and improve the coverage of the “Box” repeats in the RepeatsDB database, as well as improve the identification and study of these proteins.

ACKNOWLEDGMENTS AND FUNDING

This work was supported by 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

Acharya, S., Dahal, A., & Bhattarai, H. K. (2021). Evolution and origin of sliding clamp in bacteria, archaea and eukarya. *PloS One*, *16*(8), e0241093. https://doi.org/10.1371/journal.pone.0241093

Appleton, B. A., Loregian, A., Filman, D. J., Coen, D. M., & Hogle, J. M. (2004). The cytomegalovirus DNA polymerase subunit UL44 forms a C clamp-shaped dimer. *Molecular Cell*, *15*(2), 233–244. https://doi.org/10.1016/j.molcel.2004.06.018

Baltz, J. L., Filman, D. J., Ciustea, M., Silverman, J. E. Y., Lautenschlager, C. L., Coen, D. M., Ricciardi, R. P., & Hogle, J. M. (2009). The crystal structure of PF-8, the DNA polymerase accessory subunit from Kaposi's sarcoma-associated herpesvirus. *Journal of Virology*, *83*(23), 12215–12228. https://doi.org/10.1128/JVI.01158-09

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A., & Sillitoe, I. (2017). CATH: An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, *45*(D1), D289–D295. https://doi.org/10.1093/nar/gkw1098

Delucchi, M., Schaper, E., Sachenkova, O., Elofsson, A., & Anisimova, M. (2020). A New Census of Protein Tandem Repeats and Their Relationship with Intrinsic Disorder. *Genes*, *11*(4), 407. https://doi.org/10.3390/genes11040407

Di Domenico, T., Potenza, E., Walsh, I., Parra, R. G., Giollo, M., Minervini, G., Piovesan, D., Ihsan, A., Ferrari, C., Kajava, A. V., & Tosatto, S. C. E. (2014). RepeatsDB: A database

of tandem repeat protein structures. *Nucleic Acids Research*, *42*(Database issue), D352-357. https://doi.org/10.1093/nar/gkt1175

El Omari, K., Iourin, O., Kadlec, J., Fearn, R., Hall, D. R., Harlos, K., Grimes, J. M., & Stuart, D. I. (2014). Pushing the limits of sulfur SAD phasing: De novo structure solution of the N-terminal domain of the ectodomain of HCV E1. *Acta Crystallographica. Section D, Biological Crystallography*, *70*(Pt 8), 2197–2203. https://doi.org/10.1107/S139900471401339X

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231.

Gulbis, J. M., Kelman, Z., Hurwitz, J., O'Donnell, M., & Kuriyan, J. (1996). Structure of the C-Terminal Region of p21WAF1/CIP1 Complexed with Human PCNA. *Cell*, *87*(2), 297–306. https://doi.org/10.1016/S0092-8674(00)81347-1

Kajava, A. V. (2012). Tandem repeats in proteins: From sequence to structure. *Journal of Structural Biology*, *179*(3), 279–288. https://doi.org/10.1016/j.jsb.2011.08.009

Kajava, A. V., & Tosatto, S. C. E. (2018). Editorial for special issue "Proteins with tandem repeats: Sequences, structures and functions"☆. *Journal of Structural Biology*, *201*(2), 86–87. https://doi.org/10.1016/j.jsb.2017.12.011

Kawai, A., Hashimoto, H., Higuchi, S., Tsunoda, M., Sato, M., Nakamura, K. T., & Miyamoto, S. (2011). A novel heterotetrameric structure of the crenarchaeal PCNA2-PCNA3 complex. *Journal of Structural Biology*, *174*(3), 443–450. https://doi.org/10.1016/j.jsb.2011.02.006

Kong, X. P., Onrust, R., O'Donnell, M., & Kuriyan, J. (1992). Three-dimensional structure of the beta subunit of E. coli DNA polymerase III holoenzyme: A sliding DNA clamp. *Cell*, *69*(3), 425–437. https://doi.org/10.1016/0092-8674(92)90445-i

Lehman, I. R., Bessman, M. J., Simms, E. S., & Kornberg, A. (1958). Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme

from Escherichia coli. *The Journal of Biological Chemistry*, *233*(1), 163–170.

Mace, D. C., & Alberts, B. M. (1984). T4 DNA polymerase. Rates and processivity on single-stranded DNA templates. *Journal of Molecular Biology*, *177*(2), 295–311. https://doi.org/10.1016/0022-2836(84)90458-3

Marcotte, E. M., Pellegrini, M., Yeates, T. O., & Eisenberg, D. (1999). A census of protein repeats. *Journal of Molecular Biology*, *293*(1), 151–160. 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*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Miyachi, K., Fritzler, M. J., & Tan, E. M. (1978). Autoantibody to a nuclear antigen in proliferating cells. *Journal of Immunology (Baltimore, Md.: 1950)*, *121*(6), 2228–2234.

Moore, A. D., Björklund, A. K., Ekman, D., Bornberg-Bauer, E., & Elofsson, A. (2008). Arrangements in the modular evolution of proteins. *Trends in Biochemical Sciences*, *33*(9), 444–451. https://doi.org/10.1016/j.tibs.2008.05.008

Murayama, K., Nakayama, S., Kato-Murayama, M., Akasaka, R., Ohbayashi, N., Kamewari-Hayami, Y., Terada, T., Shirouzu, M., Tsurumi, T., & Yokoyama, S. (2009). Crystal structure of epstein-barr virus DNA polymerase processivity factor BMRF1. *The Journal of Biological Chemistry*, *284*(51), 35896–35905. https://doi.org/10.1074/jbc.M109.051581

Oakley, A. J. (2016). Dynamics of Open DNA Sliding Clamps. *PloS One*, *11*(5), e0154899. https://doi.org/10.1371/journal.pone.0154899

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*, *49*(D1), D452–D457. https://doi.org/10.1093/nar/gkaa1097

Paladin, L., Necci, M., Piovesan, D., Mier, P., Andrade-Navarro, M. A., & Tosatto, S. C. E. (2020). A novel approach to investigate the evolution of structured tandem repeat protein families by exon duplication. *Journal of Structural Biology*, *212*(2), 107608. https://doi.org/10.1016/j.jsb.2020.107608

Parrilla-Castellar, E. R., Arlander, S. J. H., & Karnitz, L. (2004). Dial 9-1-1 for DNA damage: The Rad9-Hus1-Rad1 (9-1-1) clamp complex. *DNA Repair*, *3*(8–9), 1009–1014. https://doi.org/10.1016/j.dnarep.2004.03.032

Pascal, J. M., Tsodikov, O. V., Hura, G. L., Song, W., Cotner, E. A., Classen, S., Tomkinson, A. E., Tainer, J. A., & Ellenberger, T. (2006). A flexible interface between DNA ligase and PCNA supports conformational switching and efficient ligation of DNA. *Molecular Cell*, *24*(2), 279–291. https://doi.org/10.1016/j.molcel.2006.08.015

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*, *12*(85), 2825–2830.

Piovesan, D., Tabaro, F., Paladin, L., Necci, M., Micetic, I., Camilloni, C., Davey, N., Dosztányi, Z., Mészáros, B., Monzon, A. M., Parisi, G., Schad, E., Sormanni, P., Tompa, P., Vendruscolo, M., Vranken, W. F., & Tosatto, S. C. E. (2018). MobiDB 3.0: More annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Research*, *46*(D1), D471–D476. https://doi.org/10.1093/nar/gkx1071

Randell, J. C. W., & Coen, D. M. (2004). The herpes simplex virus processivity factor, UL42, binds DNA as a monomer. *Journal of Molecular Biology*, *335*(2), 409–413. https://doi.org/10.1016/j.jmb.2003.10.064

Rasheed, N., Lima, T. B., Mercaldi, G. F., Nascimento, A. F. Z., Silva, A. L. S., Righetto, G. L., Bar-Peled, L., Shen, K., Sabatini, D. M., Gozzo, F. C., Aparicio, R., & Smetana, J. H. C. (2019). C7orf59/LAMTOR4 phosphorylation and structural flexibility modulate Ragulator assembly. *FEBS Open Bio*, *9*(9), 1589–1602. https://doi.org/10.1002/2211-5463.12700

Sakurai, S., Kitano, K., Yamaguchi, H., Hamada, K., Okada, K., Fukuda, K., Uchida, M., Ohtsuka, E., Morioka, H., & Hakoshima, T. (2005). Structural basis for recruitment of human flap endonuclease 1 to PCNA. *The EMBO Journal*, *24*(4), 683–693. https://doi.org/10.1038/sj.emboj.7600519

Schaper, E., & Anisimova, M. (2015). The evolution and function of protein tandem repeats in plants. *New Phytologist*, *206*(1), 397–410. https://doi.org/10.1111/nph.13184

Schrödinger, LLC. (2015). *The PyMOL Molecular Graphics System, Version 1.8*.

The UniProt Consortium. (2023). UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, *51*(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2022). *Fast and accurate protein structure search with Foldseek* [Preprint]. Bioinformatics. https://doi.org/10.1101/2022.02.07.479398

Venclovas, C., & Thelen, M. P. (2000). Structure-based predictions of Rad1, Rad9, Hus1 and Rad17 participation in sliding clamp and clamp-loading complexes. *Nucleic Acids Research*, *28*(13), 2481–2493. https://doi.org/10.1093/nar/28.13.2481

Warbrick, E., Lane, D. P., Glover, D. M., & Cox, L. S. (1995). A small peptide inhibitor of DNA replication defines the site of interaction between the cyclin-dependent kinase inhibitor p21WAF1 and proliferating cell nuclear antigen. *Current Biology: CB*, *5*(3), 275–282. https://doi.org/10.1016/s0960-9822(95)00058-3

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191. https://doi.org/10.1093/bioinformatics/btp033

Zhang, Y., & Skolnick, J. (2005). TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309. https://doi.org/10.1093/nar/gki524

Zhuang, Z., & Ai, Y. (2010). Processivity factor of DNA polymerase and its expanding role in normal and translesion DNA synthesis. *Biochimica Et Biophysica Acta*, *1804*(5), 1081–1093. https://doi.org/10.1016/j.bbapap.2009.06.018

SUPPLEMENTARY MATERIAL

Supplementary Table 1. ***DNA clamps with experimental structures.*** *In bold the structures analyzed in this work.*

Domain of Life	UniProt ID	PDB ID and Chain
Archaea	D0VWY8	3ifvA, 3ifvB, 3ifvC, **3hi8A**, 3hi8B, 3hi8C, 3hi8D, 3hi8E, 3hi8F
	C5A5N6	**5a6dA**, 5a6dB, 7n5iB, 7n5iC, 7n5iD, 7n5iA, 7n5jB, 7n5jC, 7n5jD, 7n5jA, 7n5kA, 7n5kC, 7n5kB, 7n5kD, 7n5lA, 7n5lB, 7n5lC, 7n5lD, 7n5mB, 7n5mA, 7n5nB, 7n5nA
	Q9UYX8	6t8hD, **6t8hE**, 6t7xA, 6t7yA
	Q9YFT8	**6aigA**
	Q975N2	**1ud9A**, 1ud9B, 1ud9C, 1ud9D
	Q5JF32	6kncE, 6kncC, 6kncD, 6knbC, 6knbD, 6knbE, 5da7A, 5daiA, **3lx1A**
	Q97Z84	7rpxB, 2io4D, 2io4B, 7rpwB, **2hiiB**, 7rpoB, 2hikB, 2ix2B, 2ntiH, 2ntiE, 2ntiB, 2izoB
	P57766	7rpxA, 2io4C, 2io4A, 7rpwA, **2hiiA**, 7rpoA, 2hikA, 2ix2A, 2ntiG, 2ntiD, 2ntiA, 2izoC
	P57765	7rpoC, 7rpxC, 7rpwC, 2ijxC, 2ijxD, 2ijxA, 2ijxB, 2hiiC, **2hikC**, 2ix2C, 2ntiF, 2ntiC, 2ntiI
	O29912	1rxmA, **1rwzA**, 3p83A, 3p83B, 3p83C, 1rxzA
	O73947	1isqA, 5aujA, **1iz4A**, 3a2fB, 1iz5A, 1iz5B, 1ge8A

	Q973F5	3aizC, 3aizD, **3aixA**
	Q5JFD3	**3lx2A**, 3lx2B, 3lx2C
	Q975M2	3aizA, 3aizB, **3aixB**
Eukarya	Q99638	3ggrA, 3a1jA, 6hm5B, 7z6hA, **3g65A**, 6j8yA
	O60671	3ggrC, **3a1jC**, 6j8yC, 7z6hB, 3g65B
	O60921	3ggrB, **3a1jB**, 6j8yB, 7z6hC, 3g65C
	B5TV91	**7wp3E**, 7wp3A, 7wp3B, 7wp3C, 7wp3D, 7wp3F, 5hacD, 5hacA, 5hacB, 5hacC, 5hacE, 5hacF, 5h0tA, 5h0tD, 5h0tF, 5h0tE, 5h0tB, 5h0tC, 6k2mF, 6k2mE, 6k2mA, 6k2mD, 6k2mB, 6k2mC, 6j0jB, 6j0jA, 6j0jD, 6j0jE, 6j0jC, 6j0jF, 5cfkA, 5cfkB, 5cfkC, 5cfkD, 5cfkE, 5cfkF
	A6ZL36	**3k4xA**
	A0A0J5SJF1	**5tupA**, 5tupB, 5tupC
	C4M9R9	**3p91A**
	G1E6N7	**4cs5A**, 4cs5B, 4cs5C
	Q9M7Q7	6o09C, 6o09A, 6o09K, **2zvvA**, 2zvvB
	P17917	**4hk1A**
	P12004	4rjfA, 4rjfC, 4rjfE, 5mloE, 5mloA, 5mloC, 6s1nE, 6s1nF, 6s1nG, 5mavE, 5mavA, 5mavC, 5mavB, 5mavF, 5mavD, 6cbiA, 6cbiC, 6cbiE, 6cbiB, 6cbiD, 6cbiF, 6qcgA, 6qcgC, 6qcgD, 6qcgB, 6qcgF, 6qcgE, 6hvoC, 6hvoA, 6hvoB, 3ja9C, 3ja9A, 3ja9B, 5ycoC, 5ycoA, 5ycoB, 5ycoD, 6k3aE, 6k3aA, 6k3aC, 6gwsA, 6gwsB, 6gwsC, 4d2gA, 4d2gB, 4d2gC, 6fcmA, 6fcmC, 6fcnA, 6fcnC, 5e0vA, 5e0vB, 5e0tA, 5e0tB, 5e0tC, 5e0uA, 5e0uB, 5e0uC, 7m5nA, 7m5nB, 7m5nC, 7m5mC, 7m5mA, 7m5mB, 7m5lC, 7m5lA, 7m5lB, 7nv1B, 7nv1C, 7nv0D, 7nv0B, 7nv0C, **4ztdA**, 4ztdB, 4ztdC, 6tnyE, 6tnyF, 6tnyG, 6tnzE, 6tnzF, 6tnzG, 1u76A, 1u76C, 1u76E, 1axcE, 1axcA, 1axcC, 5momA, 5momB, 5momC, 6gisC, 6gisA, 6gisB, 3tblA, 3tblB, 3tblC, 5mlwA, 5mlwC, 5mlwE, 6qc0A, 6qc0C, 3vkxA, 7efaA, 1w60A, 1w60B, 5iy4A, 5iy4C, 5iy4E, 6ehtA, 6ehtC, 6ehtB, 7kq1E, 7kq1A, 7kq1C, 7kq0A, 7kq0C, 7kq0E, 1u7bA, 1vymA, 1vymB, 1vymC, 1vyjK, 1vyjA, 1vyjC, 1vyjE, 1vyjG, 1vyjI, 6s1oG, 6s1oE, 6s1oF, 6s1mE, 6s1mF, 6s1mG, 3wgwB, 3wgwA, 6vvoF, 6vvoG, 6vvoH, 2zvkC, 2zvkA, 3p87A, 3p87C, 3p87E, 2zvlA, 2zvlC, 2zvlE, 2zvmA, 2zvmC

	Q03392	**6qh1A**, 6qh1B, 6qh1C
	P15873	8dqzF, 8dqzG, 8dqzH, 8dqxF, 8dqxG, 8dqxH, 3v61B, 3v60B, 4l60A, 6cx2A, 5t9dA, 5t9dB, 5t9dC, 7tfjF, 7tfjG, 7tfjH, 7tfiF, 7tfiG, 7tfiH, 7tfhH, 7tfhG, 7tfhF, 3f1wA, 6e49C, 6e49A, 6e49B, 4yhrA, 6cx3A, 1sxjF, 1sxjG, 1sxjH, 7tibF, 7tibG, 7tibH, 7ticF, 7ticG, 7ticH, 3v62B, 3v62E, 8dr4F, 8dr4G, 8dr4H, 7tidF, 7tidG, 7tidH, 7kc0G, 8dr3F, 8dr3G, 8dr3H, 5jneD, 5jneH, 3gpnA, 3gpmA, 7u1pF, 7u1pG, 7u1pH, 7tkuG, 7tkuH, 7tkuF, 3l0wA, 3l0wB, 3pgeB, 3pgeA, 6cx4A, 4l6pA, 4l6pB, 4l6pC, 5v7kA, 5v7lA, 5v7mA, 6d0rA, 6d0qA, 6w9wA, 7thvF, 7thvG, 7thvH, 3l10B, 3l10A, 7u1aF, 7u1aG, 7u1aH, 2od8A, 8dr6F, 8dr6G, 8dr6H, 3l0xA, 3l0xB, 7ti8F, 7ti8G, 7ti8H, 5zutA, 8dr1F, 8dr1G, 8dr1H, 1plrA, **1plqA**, 6wacA , 7u19F, 7u19G, 7u19H, 7thjF, 7thjG, 7thjH,
	Q5AMN0	**7bupA**, 7bupB, 7bupC
	Q7SF71	**7ep8A**
	G0SF70	**7o1eA**, 7o1fB, 7o1fC, 7o1fD, 7o1fE, 7o1fF, 7o1fA
	Q9ZW35	2zvwC, 2zvwD, 2zvwE, 2zvwF, 2zvwG, 2zvwH, **2zvwA**, 2zvwB
	Q08949	**7sh2H**, 7sgzH, 7stbG, 7st9G
	P48581	**7sh2G**, 7sgzG, 7stbF, 7st9F
	Q02574	**7sh2F**, 7stbH, 7st9H, 7sgzF
Bacteria	J1IY24	6ptrA, 6degB, **6degA**
	Q1R4N6	**4k74A**, 4k74B
	G8LES0	6amqA, 6amqB, **6amqC**, 6amqD
	V5V7W3	6ap4N, 6ap4O, 6ap4P, 6ap4F, **6ap4A**, 6ap4B, 6ap4C, 6ap4D, 6ap4E, 6ap4G, 6ap4H, 6ap4I, 6ap4J, 6ap4K, 6ap4L, 6ap4M
	P9WNU0	6fvnA, 6fvnC, **4tr7A**, 4tr7B, 6fvoD, 6fvoB, 6fvoC, 6fvoA
	P9WNU1	3rb9A, 3rb9B, 3p16D, **3p16A**, 3p16B, 3p16E, 3p16C, 3p16F, 5agvA, 5agvB, 5aguB, 5aguA
	A0QND6	**5ah2A**, 5ah2B, 5ah2C, 5ah2D, 5ah4A, 5ah4B
	Q92I37	5w7zA, 5w7zB, **6dm6A**, 6dm6B
	Q68WW0	**6d46A**, 6djkA
	A0A0H3AWV3	**6ptvA**, 6ptvB, 6ptvD, 6dlkA, 6dlkB

	Q9EVR1	**2avtA**, 2avtB
	O06672	**2awaA**, 2awaB, 2awaC, 2awaD
	B2HI47	**6d47A**, 6dlyB, 6dlyA
	C4Z938	3t0pB, **3t0pA**
	P05649	**4tr6A**, 4tr6B, 6e8dC, 6e8dD, 6e8dB, 6e8dA
	P33761	**6dj8A**, 6dj8B
	P0CAU5	5wceB, **5wceA**, 6izoA, 6izoB, 6jirB, 6jirA
	P0A990	**5x06A**, 5x06B, 5x06C, 5x06D, 6fvmA, 6fvmB
	P0A988	7azfB, 7azfC, 7azfD, 7azfA, 7azgH, 7azgA, 7azgG, 7azgB, 7azgC, 7azgD, 7azgE, 7azgF, 7azdA, 7azdB, 7azdC, 7azdD, 7azeA, 7azeB, 7azcA, 7azcB, 7azcC, 7azcD, 3qsbA, 3qsbB, 3q4jA, 3q4jB, 3q4jE, 3q4jC, 3q4jD, 3q4jF, 2xurB, 2xurA, 3q4lA, 3q4lB, 7azkC, 7azkA, 3f1vB, 3f1vA, 3pweA, 3pweB, 1jqjA, 1jqjB, 4n99B, 4n99A, 6e8eA, 6e8eB, 4n98A, 4n98B, 1mmiB, 1mmiA, 4k3oA, 4k3oB, 4k3lA, 4k3lB, 4k3mA, 4k3mB, 4k3kA, 4k3kB, 3d1fA, 3d1fB, 3d1gA, 3d1gB, 3d1eB, 3d1eA, 4k3rB, 4k3rA,4mjqA, 4mjqB, 4k3pB, 4k3pA, 4k3qA, 4k3qB, 4n95B, 4n95A, 4n94B, 4n94A, 4n97B, 4n97A, 2polA, 2polB, 5m1sB, 5m1sC, 3q4kA, 3q4kB, 6fvlD, 6fvlC, 6fvlA, 6fvlB, 4k3sB, 4k3sA, 7az6A, 7az7A, 7az5D, 7az5C, 7az5A, 7az5B, 7azlD, 7azlA, 7azlB, 7azlC, 4mjrA, 4mjrB, 7az8A, 7az8B, 4mjpA, 4mjpB, 1jqlA, 4n9aA, 4n9aB, 1ok7A, 1ok7B, 5fkuC, 5fkuB, 5fkvB, 5fkvC, 5fkwB, 5fkwC, 1unnB, 1unnA, **4n96A**, 4n96B
	O25242	5g4qA, 5g4qB, 5g48A, 5g48B, 5fxtA, **4rkiA**, 5fveA, 4s3iA, 4s3iB, 5frqD, 5frqA, 5frqB, 5frqC
	Q9I7C4	**4tr8A**, 4tr8B, 6amsD, 6amsA, 6amsB, 6amsC, 6pthA, 4tszK, 4tszI, 4tszO, 4tszM, 4tszA, 4tszC, 4tszE, 4tszG
	Q1RIS7	**6manA**, 6manB
	Q9RYE8	**4trtA**, 4trtB
	Q9WYA0	**1vpkA**
	P0A024	**7evpA**
	B2FT80	**7rzmA**
	O80164	1b8hA, 1b8hB, 1b8hC, **1b77A**, 1b77B, 1b77C

Viruses	P04525	3u60H, 3u61G, 3u61H, 3u61F, 3u5zP, 3u5zG, 3u5zH, 3u5zF, 3u5zQ, 3u5zR, 6drtC, 6drtA, 6drtB, **1czdA**, 1czdB, 1czdC
	P03191	**2z0lA**, 2z0lB, 2z0lC, 2z0lD, 2z0lE, 2z0lF, 2z0lG, 2z0lH
	Q77ZG5	3i2mX, **3hslX**
	P16790	1yypA, 5ixaB, 5ixaA, 5iwdA, **1t6lA**
	P10226	1dmlC, 1dmlE, **1dmlA**, 1dmlG

Supplementary Figure S1. Structural similarity and sequence identity comparison between the different Pfam domains present in DNA clamps of Archaea, Eukarya and Bacteria. Structural similarity is given by TM-Score represented as a percentage.