Published online 13 February 2019

Nucleic Acids Research, 2019, Vol. 1, No. 1 **1–9**
doi:10.1093/nar/gkn000

A new Census of Protein Tandem Repeats

Matteo Delucchi [1] **and Maria Anisinova** [2, *]

[1], [2] ZHAW School of Life Sciences and Facility Management, Institute for Applied Simulations, Einsiedlerstrasse 31, 8820 Waedenswil, Switzerland

Received January 1, 2009; Revised February 1, 2009; Accepted March 1, 2009

ABSTRACT

We analyze systematically and with state of the art methods all known protein sequences for tandem repeats. From all curated proteins of all domains of life, 50.9% contained at least one tandem repeat. Eukaryotic proteins tend to have more TRs than prokaryotic which could be explained by the complexity of the respective organisms. A postive linear correlation between the amount of TR units and protein length could be detected. This correlation becomes weaker with increasing TR unit size. TRs often didn't appear alone in the same protein. 43% of eukaryotic proteins have even more than four distinct TR per protein. We further saw that small TRs appear more frequently and we showed that TRs are non-uniform distributed across the protein sequence. They are mostly located towards the ends. We provide insights in the origin and function of proteomic tandem repeats accross all kingdoms of life.

INTRODUCTION

The continued progress in genomics demands better classification and understanding of genomics sequences, their evolution and function across the tree of life. Proteins indisputably remain at the heart of the molecular machinery performing a multitude of essential functions.

Genomic versus Proteomic Tandem Repeats

According to most recent estimates a substantial amount of proteins contain adjacently repeated amino acid (AA) sequence patterns, known as tandem repeats (TRs). Analogously to repeated sequence patterns in DNA, they are called *homogeneous* (homo-) or *heterogeneous* (hetero-) repeats for consisting of identical units or mixed units respectively (1) and can be either classified as *direct repeats* for a head-to-tail or *inverted repeats* for a head-to-head orientation (2). TRs are described by a certain length of their repeating motif (unit length), their number of repeated units (size) and the similarity among their units (12).

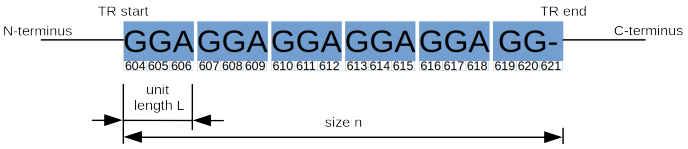


Figure 1. A sketch of a tandem repeat with it's descriptors. This micro TR with the ID A7TKR8 and a size of 6 units, each with a unit length of 3 amino acids shows a head-to-head orientation and consists of mixed units - a direct- and heterorepeat.

Depending on their size, DNA TRs are classified into microsatellites (1-8 nucleotides) and minisatellites (>9 nucleotides) (2). They are either perfect or imperfect repeats depending on whether they are exact copies of one another or deviate by more than one base pair (3). We use a similar nomenclature for protein TRs: Protein TRs with a length L of 1 amino acid are herein called homo tandem repeats (homo TRs), protein TRs with $1 < L \leq 3$ amino acids are called micro tandem repeats (micro TRs), small tandem repeats (small TRs) for a length of $4 < L \leq 15$ and domain tandem repeats (domain TRs) for protein TRs of a length of ≥ 15 amino acids. In figure 1 a graphical representation of the descriptors of a protein TR sequence is shown.

In the human proteome TRs are with more than 55% abundancy of repetitive elements (27) highly represented and display an impressive variability of sizes, structures and functions (4, 5). Proteins containing TRs have enhanced binding properties (21) and are known to have associations with immunity related functions (22, 23) and diseases such as amyotrophic lateral sclerosis (ALS), myotonic dystrophy (DM), dentatorubral-pallidoluysian atrophy (DRPLA), frontotemporal dementia, fragile X syndrome (FXS), fragile X tremor-ataxia syndrome (FXTAS), Huntington disease, spinobulbar muscular atrophy (SBMA) and spinocerebellar ataxia (SCA) which are all caused through tandem repeat disorders (TRD) (28).

*Dr. Maria Anisimova, ZHAW School of Life Sciences and Facility Management, Institute for Applied Simulations, Computational Genomics, Tel: +41 (0)58 9345882; Email: maria.anisimova@zhaw.ch

Genomic Tandem Repeat Origination & Evolution

Similarity between the TR units fades with time since TRs can evolve either during meiosis or mitosis by processes such as duplication and loss of TR units, recombination, replication slippage and gene conversion which all can cause changes to their unit similarity and length (6). This evolution in TR units makes them a rich source for genetic variability by providing a wide range of possible genotypes at a given locus (7). Therefore, they are prone sites for selection on long evolutionary scales as well as on a somatic level. The occurrence of mutations in TR within protein coding genes, can alter the structure and therefore likely the function of the affected protein too. Since non-coding regions play crucial roles in gene regulation, transcription, and translation, the proteins concerned are also likely to be affected by TR-mutations occurring in non-coding sequences. While the biological mechanisms generating TRs are not well understood, evidence suggests that natural selection contributes to shaping TR evolution (5), and that TR expansion is linked to the origin of novel genes (59) TRs have been successfully exploited in bioengineering due to their "design-ability" (8).

A Comprehensive Analysis of Proteomic Tandem Repeats

Despite much interest (9), the most recent and commonly cited census of protein TRs summarizing repeats from UniProtKB/Swiss-Prot protein knowledgebase (26) dates back two decades (10). Since then the number of proteins in the curated protein databank SwissProt has grown more than seven fold (supplementary figure S1). Equally, a multitude of new methods were developed for the prediction and analysis of TRs (11, 24, 25). In particular, due to striking differences in TR predictor properties, a new statistical framework and a meta-prediction approach was proposed in order to increase the accuracy and power of the TR annotation (11, 12).
Here we apply this recent methodology to characterize the distribution of protein TRs as found in the up-to date SwissProt protein knowledgebase (13, 26). Our TR annotation for each protein includes the TR region start, end, minimal repeated unit length, among unit divergence and TR unit alignments. This allows our study to provide an unprecedented detail of the universe of protein TRs.
To our knowledge, no recent study connected the information of viral proteomic TRs to their hosts TRs. Viral genomes are generally relatively small which demands for an optimal coding capacity. Furthermore, they lack their own translational apparatus and depend completely on their hosts protein sythesis machinery (41, 42). This makes them an interesting subject of research and therefore we provide a large-scale virus TR study where we compare the TR distribution of the viral proteome and their hosts proteome.
Further, proteins with TRs tend to be enriched with intrinsic protein disorder (IDP) (17), and vice versa (14). IDPs cover multiple three dimensional states of proteins leading to different functionalities (32). Both TR and intrinsic disordered regions (IDR) also tend to be overrepresented in the hubs of protein-protein interaction networks (15). While the relationship between these non-globular protein features has been observed, the biological reasons are not well understood. TRs often fold into specific structures, such as solenoids, or have "beads on a string" organizations (16). But there is undoubtedly a class of protein TRs strongly associated with unstructured regions (14, 17). Several studies have shown that compositionally biased, low complexity regions, often found in IDPs evolve rapidly, including recombinatorial repeat expansion events (18, 19). Others in contrast observed that the association between repeat enrichment and protein disorder is not as clear (20). In order to systematically characterize and explore the enigmatic connection between TRs with IDP, we also use state of the art methods to annotate each protein with IDP regions and summarize the distribution of the overlap of TR and intrinsic disordered regions over all kingdoms of life. We distinguish four types of overlap (suppl. figure S2). We call the overlap a *tail-overlap* where intrinsic disorder begins within the TR-sequence and finishes after the TR-region. In contrast, we call it *head-overlap* if the IDR begins before the TR-sequence and finishes within. If the IDR lies completely within a TR sequence, we call it *Disorder-in-TR* and *TR-in-Disorder-overlap* if the TR-region lies within the IDR. This characterisation of IDR-overlaps with TRs offers an unprecedented detailed view of the interplay of TRs with IDR.

RESULTS

Exhaustive annotation of protein TRs in the entire UniProtKB/Swiss-Prot was done using a meta-prediction approach based on both de novo and profile-based methods followed by filtering of false positives and redundancies. The pipeline was implemented in Python using TRAL (12); see Methods for details. Structural and biochemical properties of TRs can be extremely diverse depending on the length and the composition of their minimal repeating unit. We studied TR properties in four categories defined according to TR unit length L: (1) homorepeats, (2) microrepeats, (3) small repeats, and (4) domain repeats.

Impressive numbers of TR annotations were predicted; their distributions with respect to protein length and repeat number are summarized by Superkingdoms in table 1 and in the suppl. figures S3a and S3b.

TRs are Abundant in Proteins of all Domains of Life

Overall, 50.9% of all eukaryotic proteins contained at least one TR. In *Homo sapiens* (Human), 68.8% of all proteins contain TRs. Similar to *Mus musculus* with 61.9% and *Drosophila melanogaster* with 60.8%. In contrast stands *Escherichia coli* with 28%. Interestingly, 43.6% of viral proteins contained TRs, almost as frequently as in Eukaryotes. In comparison, fewer prokaryotic proteins contained TR, but nevertheless >30% for both bacterial and archaeal proteins.

SwissProt Census				
of all proteins	Archaea	Bacteria	Eukaryota	Viruses
TR count	6420	103842	92472	7237
TR fraction	0.331	0.312	0.509	0.436
homo TR fraction	0.006	0.006	0.086	0.029
micro TR fraction	0.117	0.109	0.245	0.191
small TR fraction	0.217	0.208	0.328	0.300
domain TR fraction	0.051	0.049	0.143	0.069
mean prot. sequence length	288	313	436	451
prot. count	19370	332327	181814	16605
of proteins containing TRs				
homo TR fraction	0.019	0.019	0.169	0.067
micro TR fraction	0.354	0.350	0.482	0.438
small TR fraction	0.656	0.667	0.644	0.689
domain TR fraction	0.154	0.157	0.281	0.158
mean prot. sequence length	355	404	572	644
prot. count	6420	103842	92472	7237

Table 1. Swissprot entries by kingdoms for all proteins and for proteins that contain TR. Over all proteins, Bacteria has the most entries but Eukaryota the biggest fraction of proteins with TRs. Viruses tend to have the longest protein sequences - with or without TRs; followed by eukaryotic and prokaryotic proteins. In general, small TR prevail over the other types.

Proteins containing homo TRs have generally TRs mostly of small size (mean = 8.8 repeat units). They make up 20% of all found TRs over all Superkingdoms and 30% for Human TR. Of all the homo TRs, 91.3% are from Eukaryotic origin. We couldn't detect a protein which contains only homo TRs and no other type of repeat.

Proteins with micro TRs tend to have TRs with a mean of 7 repeat units. When we looked at proteins which contain only TRs of the type micro, the mean repeat unit number = 3. Proteins containing micro TRs make up 56% of the found TRs.

Proteins with small TRs tend to have TRs with less repetition units than those with homo TRs (mean = 6 repeat units). When we looked at proteins which contained solely TRs of the type small, the mean repeat unit number is 3. Proteins containing small TRs make up 76% of the found TRs.

Domain TRs mostly consist of few units (mean = 3.5 repeat units). A prominent exception is an extracellular matrix-binding protein (Q5HFY8, *S. aureus*) with 80 units each 97aa (PF07564) spanning 7700aa. Other exceptions of bacterial domain TRs with many units are the cell surface glycoprotein 1 of *Clostridium thermocellum* and some uncharacterized PE-PGRS family proteins of *Mycobacterium tuberculosis*. Proteins containing domain TRs make up only 30% of all found TRs.

Eukaryotic domain TRs tend to be more uniform distributed in terms of their TR size. The proteins with the most repeat units belong to mediator of RNA polymerase II transcription subunit proteins from yeast (*Eremothecium gossypii*), slime mold (*Dictyostelium discoideum*) and Human Mucin-22 protein. Figure S3a shows a peak of proteins containing many TR units. They seem to belong to collagen-like proteins of the Mimiviridae family.

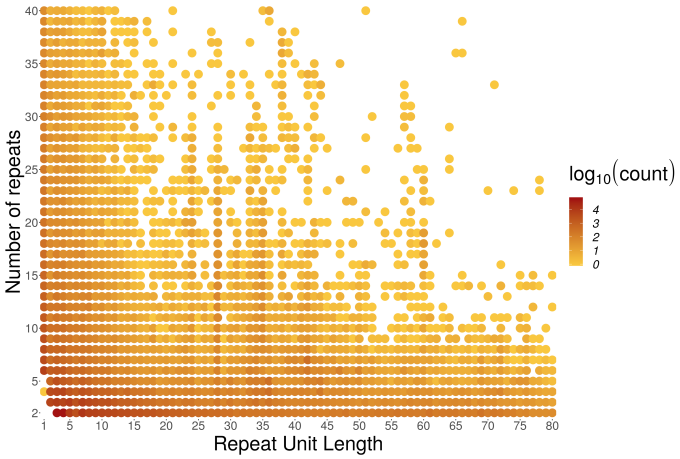


Figure 2. Distribution of TRs as a function of their repeat unit length $l_{effective}$<=80 and their number of repeat units $n_{effective}$<=40. Darker colour indicates a larger number of TRs with a specific length and number of repeats. The majority of TRs has small TR units. Yet, there is a blob of domain TRs ($25 < l_{effective} < 50$), with certain TR unit length clearly enriched (e.g., $l_{effective} = 28$, mostly zinc finger TRs.)

TRs vary in their Length and Number

In general, TRs are not homogenously distributed in terms of their unit lengths and numbers. Figure 2 reveals multiple peaks, showing that some unit lengths are particularly frequent. These peaks represent common TRs, with specific TR units used in varying number. One such example are zinc finger proteins, abundantly present as a TR in all domains of life, but also LRR and WD40-like beta propeller.

In Bacteria (*Porphyromonas gingivalis*) hemagglutinin A is known to be involved in host colonisation by adhesion to extracellular matrix proteins and is expected to be involved in periodontal diseases (29, 30). It can be seen

in the suppl. figure S3b as one of the bacterial outliers (in terms of unit length) with a unit length >450. The other two outliers belong to the Mannuronan epimerase protein of *Azotobacter vinelandii*. Eukaryotes tend to have in general the longest TR units with five particular big outliers which belong to the Anchorage 1 protein and Nesprin homolog in *Caenorhabditis elegans* and Mucin-12 and FCγBP (which has mucin-like structure (31)) in Humans.

Multiple Tandem Repeats per Protein

A substantial fraction of proteins contained more than one distinct TR region, most frequently in eukaryotic proteins (56% of all proteins with TRs), but also in viral (45.7%) and prokaryotic proteins (28.4% in Bacteria and 26.6% in Achaea). In Eukaryotes, 43% (90026 absolute count) of all proteins with TRs had 4 (or more) distinct TR regions. After them come Viruses with 28.6% followed by Bacteria 9.1% and Archea with 8.0% having ≥4 distinct TR regions per protein.

In proteins which have ≥4 TRs, the TR-types are not necessarily the same. By far the most frequent TRs in proteins containing ≥4 TR regions, were small repeats (95.0% of all predicted TRs), followed by microrepeats (87.9%), and domainrepeats (47.6%) (see suppl. figure S4).

More TRs are Found in Longer Proteins

Proteins with a chloroplastic origin tend to be shorter and contain less TR than Viridiplantae proteins from mitochondrial origin. Mitochondrial proteins are in general shorter and have less TR than proteins without endosymbiontic origin. Figure 3 displays the linear relationship between mean protein length and the amount of TR. Prokaryotic proteins cluster with their protein length and TR content in the same range as chloroplastic proteins. It seems that TRs are increasingly abundant in increasingly complex organisms (consider also suppl. figure S5).

A different level of detail is obtained by not only considering the mean of the protein length, but by grouping the amount of TRs according to bins of protein sequence length. Figure 4 shows that in general, differences in TR distributions observed between kingdoms can be largely attributed to protein sequence length with Eukaryotes having on average longer TRs. The fraction of eukaryotic proteins with homo TRs behaves differently than in other TR-types and compared to the other superkingdoms: The amount of homo TRs in eukaryotic proteins increases exponentially whereas for the other superkingdoms, the homo TR fraction stays on a similar level by increasing sequence length (see suppl. figure S6, S7, S8, S9).

Looking specifically at proteins which contain TRs, we can see in figure 5 that on average longer protein sequences tend to have more TRs and small TR seem to be the most recurrent in all kingdoms.

Indeed, we observe a postive correlation between the protein length and the fraction of proteins with TRs across all kingdoms of life for all TR types: $R^2=0.46$, $p<0.001$ for homo TRs; $R^2=0.64$, $p<0.001$ for micro TRs; $R^2=0.88$, $p<0.001$ for small TRs, and $R^2=0.24$, $p=0.08$ for domain TRs.

The correlation is slightly weaker for the domain repeats, where factors other than protein length must contribute to explain the amount of TRs, perhaps due to differences in TR generating processes for different TR types. On the other hand, consistent with the same trend, we observed that homorepeats are particularly frequent in Eukaryotes, where proteins are on average longer. Moreover, longer homorepeats are mostly characteristic to Eukaryotic proteins. For example, this can be observed from Figure 6. PolyQ and polyN homorepeats may often be observed with >50 repetitions in Eukaryotes. The

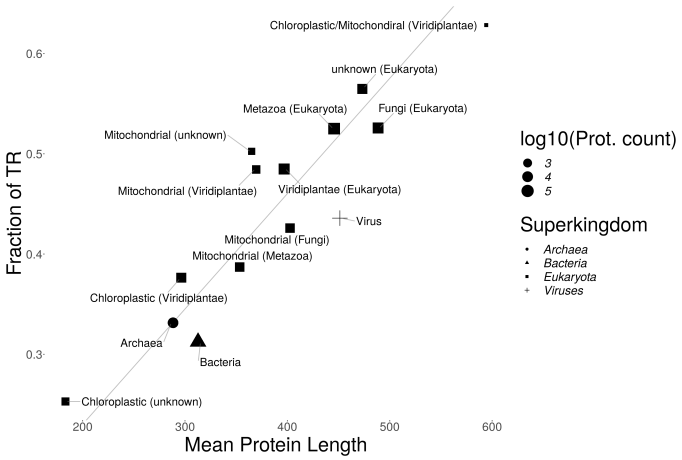


Figure 3. The fraction of proteins containing TRs over all protein entries in UniProtKB/Swiss-Prot is shown for each taxonomic domain (Superkingdom) or kingdom and displayed as function of the mean protein length and split according to the origin of the proteins. Chloroplastic proteins seem to be shorter and tend to have less TR than mitochondrial proteins. Non-mitochondrial and non-chloroplastic proteins appear to be longer and with more TRs.

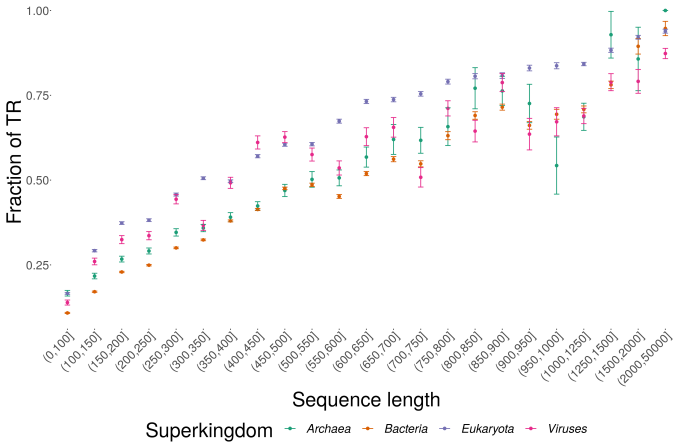


Figure 4. The fraction of proteins with TRs as a function of sequence length by kingdom resulting in a linear relationship. With Eukaryotes having on average more TRs than the other kingdoms.

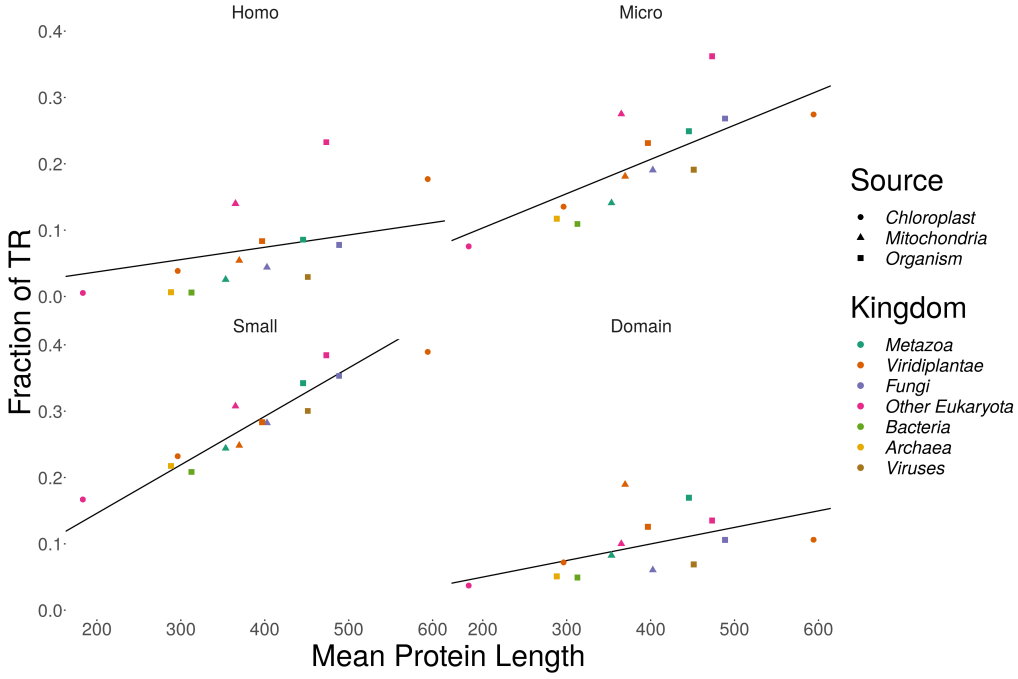


Figure 5. The amount of TRs (normalized by the amount of protein entries of the species) is displayed separately for each TR-type as a function of the mean length of the proteins. It can clearly be seen, that TRs appear mostly as small TRs. Comparing the fraction of TRs kingdom-wise, some clear tendencies can be seen for micro- and small TRs. For example, chloroplastic proteins with unknown Kingdom (better: different Kingdoms?)tend to have few TRs and short mean protein length. Where in contrast mitochondrial proteins from Viridiplantae and Fungi tend to have many TRs and long mean protein length.

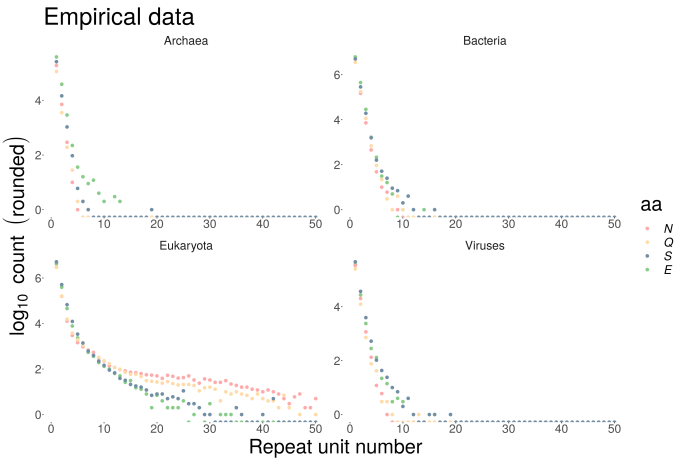


Figure 6. Count of homorepeats in Swiss-Prot in four Superkingdoms for different repeat unit number ($n <= 50$, equivalent to repeat length) for hydrophilic Asparagine (N), Glutamine (Q), Serine (S) and Glutamic acid (E). Homorepeats with large n seem to mostly pertain to the Eukaryotes.

same homorepeats display <10 repetitions for poly Q and <20 for poly N in prokaryotes and viruses. However, this large discrepancy cannot be explained purely by the length of the proteins involved.

TR-location is Biased Towards Flanks for Shorter TRs

Next, we explored where in a protein sequence tandem repeated regions tend to be found. The location within a protein was evaluated with respect to the center of a TR region and normalized by the protein length (see Methods). The observed distribution of TRs along the protein length was non-uniform and dependent on the TR unit length. Figure 7 shows the distributions of the relative positions of TRs in proteins across all different kingdoms and for different TR unit length categories.

As expected, TR relative position is shown to be prefered to the beginning of proteins. For homorepeats such tendency was particularly striking, particularly in Archaea, where most homorepeats were found in the C-terminal and domain TRs which tend to be found at the N-terminal (see suppl. figure S10). Overall, shorter TRs (homo TRs, micro TRs and small TRs) displayed stronger preferences towards both, N- and C- terminals of proteins. In particular for Eukaryotes, there was a clear correlation between the TR unit length and the location bias towards the protein flanks. Interestingly, also domain TRs in Viruses and to a smaller degree in Archaea show a similar behavior to be located towards both flanks of proteins. In eukaryotic proteins TRs were found to be overrepresented in the N-terminal protein flank, while in Archaea and Bacteria, the TR preference was towards the C-terminal.

Analogously we looked at the location of IDRs in proteins and found parallels to TRs. IDRs tend to be

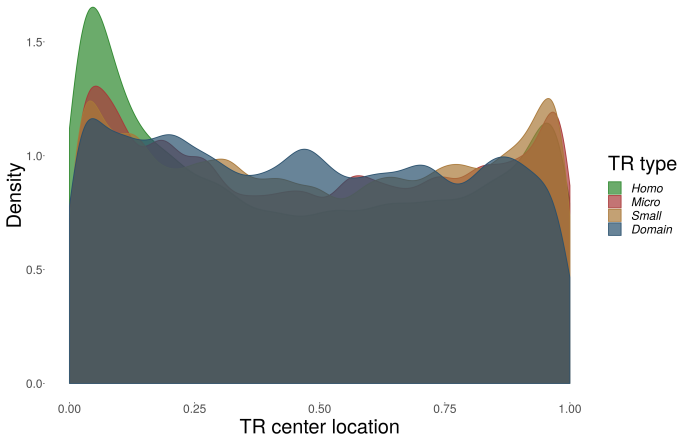


Figure 7. Density plot for the relative positions of tandem repeats (TRs) within the protein for four Superkindoms. The relative position referes with 0 to the N-terminus and with 1 to the C-terminus of a protein. Colours indicate repeat unit lengths. Interestingly, shorter TRs are biased towards the flanks of the protein.

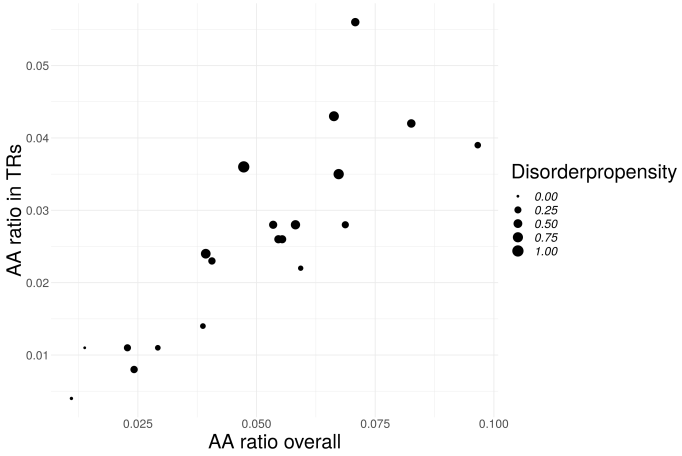


Figure 8. AA frequencies in tandem repeats against the amino acid frequency over all proteins normalized by their total amount of amino acids. An increased size of the dot corresponds to an increased disorder propensity. The distribution of the AAs in TRs corresponds to the overall distribution in proteins eliminating a distribution bias.

located towards the flanks where short IDRs prefere to be located near the N-terminal as shown in supplementary figure S11.

TRs Have a Significant Amount of Disorder Promoting Amino Acids

We further compared the abundancy of amino acids in TRs with the rest of the proteins by their disorder-promoting propensity (60).

The amino acid abundancy in TRs is linear dependent to the overall distribution in proteins as shown in figure 8. TRs can not be characterized by a certain amino acid abundancy. However, we could find a significant positive correlation (ρ= 0.71, p $<$ 0.05) of the abundancy of amino acids in TRs with the corresponding amino acid's disorder propensity (see suppl. figure S12). Where in contrast, the overall amino acid abundancy in all (incl. TR-containing) proteins showed less correlation and significance (ρ= 0.44 , p = 0.053). We couldn't detect any correlation when the TR-containing proteins were excluded from the overall fraction (ρ= -0.10, p $>$ 0.05).

Looking at the amino acids of homo TRs, grouped by their presence in disordered or ordered regions, we could see in figure 9 that amino acids which were expected to promote disorder, appear more often in homo TRs of AA with a higher disorder promoting propensity. Specifically in Eukaryotes, we could observe (see suppl. figure S13) the behaviour of homo TRs beeing relatively frequent in longer repeats compared to the other Superkingdoms. We can generally see an exponential decay of homo TRs with increasing repeat unit number except for polyhistidine (polyH), polyasparagine (polyN) and polyglutamine (polyQ). PolyN and polyQ homo TR appear to be frequently longer compared to the other residues. We further see that polyN and polyH homo TR

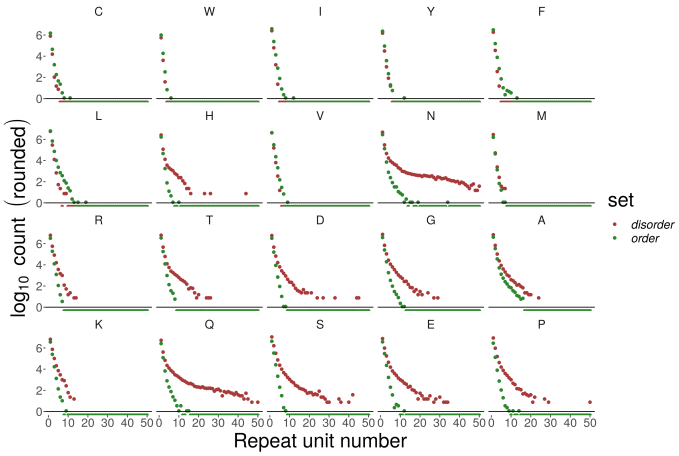


Figure 9. Empirical count of homorepeats in Swiss-Prot Eukaryotes ($n <= 50$) for ordered and disordered regions (consensus MobiDB annotations, no minimum length cut-off). Amino acids are ordered by their propensity to promote structural order. It can be seen, that polyN and polyH regions, are frequently detected within disordered protein regions eventhough their disorder promoting propensity is relatively low. Further, the frequency of polyQ and polyN residues decreases differently with increasing length, than for other AA.

which have a low disorder promoting propensity, appear relatively frequent in long repeats.

To compare the empirical observations with statistically expected observations, we estimated the number of homo TRs of a certain amino acid and repeat unit number, (which corresponds to a sequential run of success in a Bernoulli trial; see Methods) and compared this to the empirically found number of homo TRs in figure 10. Resulting in an exponential decay in frequency of homo TRs with increasing size of the homo TR.

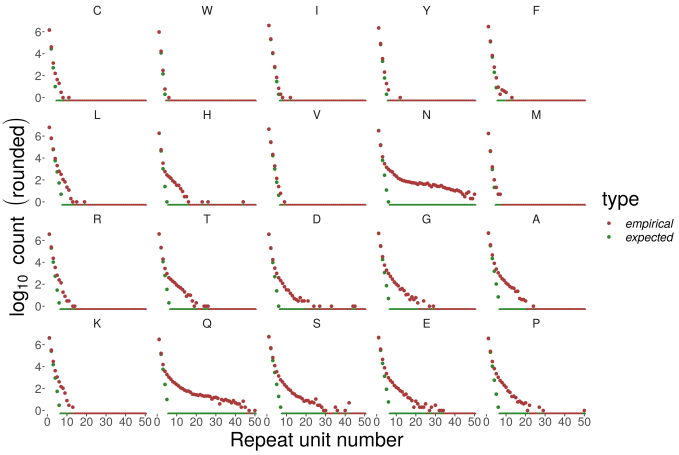


Figure 10. Empirical and expected count of homorepeats in Swiss-Prot Eukaryotes ($n <= 50$). Amino acids are ordered by their propensity to promote structural order. It can be observed, that disorder promoting residues are found more often than expected in long TRs. PolyN and polyH repeats have a low disorder propensity but show the characteristics observed in AA with high disorder propensity.

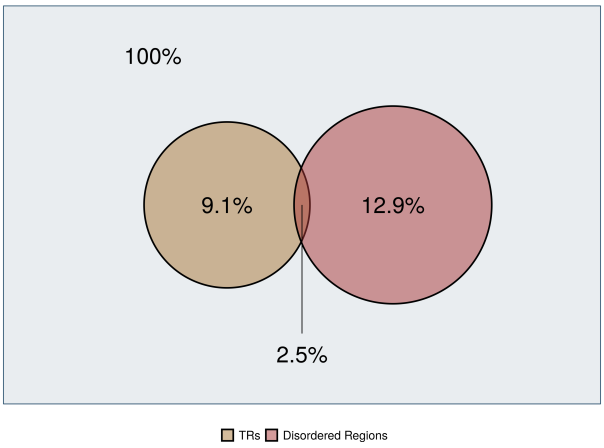


Figure 11. The areas represent the total number of amino acids of each group. Where the rectangle shows the total amount of amino acids in all proteins, the amino acids in disordered regions and the amino acids in TR-regions and their overlap. 9.1% of the length of protein sequences can be attributed to tandem repeated regions and 12.9% to disordered regions. 2.5% of the total length of protein sequences overlaps with both.

TRs are part of Intrinsic Disordered Regions

Figure 11 shows that overall a larger fraction of the total amount of protein sequence is annotated as IDR than TRs.

Intrinsic disorder regions are often found in proteins with TRs. We could see that 19.6% of the IDR overlap with tandemly repeated regions.

The largest fraction of overlap can be contributed to a complete overlap of IDRs with TRs (see figure 12). This shows that TRs are mostly a complete part of IDR.

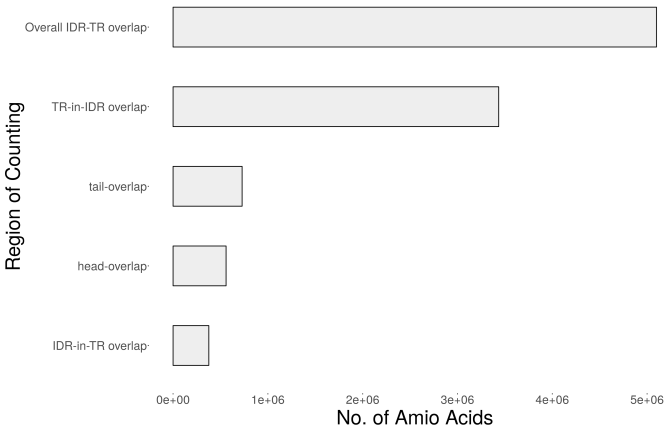


Figure 12. The absolute size of intrinsic disordered regions with TR-regions can be split in the four different kinds of overlap. We can see that TRs are most often nested within intrisic disordered regions.

TRs are Involved in Transcription Processes, Stuctural Organisation, Electron-transport & Ion-binding

From all detected TRs, 4.5% were retrieved with a PFAM model. Combining the PFAM model with the PFAM-Name we could cluster TRs in their protein families (see methods for details). Many more entries for TRs with a PFAM entry could be detected in Eukaryotes (73%) compared to the other Superkingdoms.

For each Superkingdom, we filtered the 10 most observed PFAM-families of TR containing proteins (whole list in suppl. table 1). Which resulted in the overall detection of many TRs falling within proteins which are involved in transcription. As expected, we detected in Eukaryotes many TRs in Zn-finger motifs which are responsible for adhesion to DNA, RNA and lipids. WD-40 repeats in alteri involved in transcriptional regulation. RNA recognition motifs such as the K Homology (KH) domain which binds to RNA, transcriptional repression such as the Pumilio-family found in many TR-containing proteins of fungi. Further eukaryotic proteins containing TRs appear to occur in RNA-polimerase binding and RNA-splicing by i.e. KRAB box domain as well as proteins involved in the assembly of multiprotein complex.

Additionally, many TRs were detected in proteins involved in electron-transport and ion-binding such as EF-hand domain pair and Ca^{2+}-binding EGF domain in Eukaryotes, the zinc-dependent enzyme UDP N-acetylglucosamine O-acyltransferase and the Rad50 zinc hook motif in Bacteria. Zinc-binding motifs could be detected in viral TR-containing proteins such as the zinc knuckle motif, too.

In proteins which have chloroplastic origin, we detected many TRs in NifU-like domains. They are involved in the formation of metalloclusters of nitrogenase in certain bacteria and the maturation of FeS clusters.

Proteins from mitochondrial origin showed many TRs in proteins with EF-hand domain pairs which is

found in a large family of Ca^{2+}-binding proteins and with Bacterial transferase hexapeptide which combines several transferase protein families including zinc metalloenzymes. In both, chloroplasts and mitochondria, many Pentatricopeptide repeats (PPR) could be detected. They play roles in RNA stabilisation and processing. We further found many TRs in Ankyrin repeats (especially in viruses) which are known for their diverse functions in transcription- and cell-cycle regulation, signaling and ion-transporters but also have cytoskeletal functions.

Proteins which are involved in the structural organisation of cells could be found in all Superkingdoms. Well known are the extracellular structure proteins of the collagen superfamily involved in formation of connective tissues but also leucine-rich repeats (LRR) which is unusually rich in hydrophobic amino acids forming a solenoid protein domain. They seem to provide a structural framework for the formation of protein-protein interactions. Proteins containing LRRs are involved in transcription, RNA processing, signal transduction and more.

In Eukaryotes we further found many TRs in proteins with LIM domains and tetratrico peptide repeats (TPR). LIM domains are formed by two zinc-finger domains and are involved in cytoskeletal organisation, organ development and oncogenesis. TPR and LIM motifs are both mediating protein-protein interactions. TPRs are also playing roles in cell cycle regulation, transcriptional control, protein transport, neurogenesis and protein folding.

We further saw that many TR-containing proteins of chloroplastic origin seem to contain the catalytic domain of homoserine dehydrogenase involved in the aspartate pathway which leads to the production of amino acids but also produces essential components of bacterial cell wall biosynthesis.

For the TRs in proteins which could be associated with a protein family, we calculated the TR center location (see Methods) of the most frequent PFAMs in each Superkingdom shown in figure 13.

We could see that TRs in proteins containing the Ribosomal protein L6 domain tend to locate at the same position in Archaea and Eukarya. The similar trend can be seen for Bacterial transferase hexapeptide in Archaea and Bacteria and WD-40 Beta Propeller Repeat in Bacteria and Eukaryota as well as for the TFIIB zinc-binding domain in Archaea, the zinc finger, C2H2 type of Eukaryota and zinc knuckle of Viruses. Examining closer the inter-kingdom relationship of the zinc-binding domain, we see that in Archaea, the N-terminal zinc ribbon is part of the recruitment of RNA polymerase II where a beta sheet structure of cysteine and histidine residues coordinates the zinc ion. Similarly in the viral zinc knuckle domain, a beta sheet of cysteine and histidine mediates the zinc ion. The zinc finger domain in eukaryotes is the best described one. Multiple zinc finger domains appear as tandem repeats building together the DNA binding domain of the protein by binding into the major groove of the nucleic acids double helix structure.

This does not only shows that TRs of proteins with similar function seem to cluster at the same position in the protein across all Superkingdoms but also supports the hypothesis of TRs beeing directly involved in binding activities to nucleic acids and therefore beeing involved in transcriptional regulation.

Viruses have less TRs than their host organisms

The TR-types are distributed in the similar proportion for Prokaryotes (2:4:1 for micro:short:domain) and between Eukaryota and Viruses (1:3:∼1 for micro:short:domain). Therefore we searched for further similarities in terms of TR content but not only between Eukaryotes and Viruses but more general between TR content of Viruses and their hosts.

Of all proteins, only 3% are viral. Of those viral proteins, 43.6% contain at least one TR and 58.7% have only a single TR per protein. Only a negligibly small part of them, don't have an annotated virus-host species.

Of all viral proteins (incl. those without TRs), most of them have an eukaryotic virus-host (92%) followed by bacterial virus-host (6%) and archael virus-host (2%). Most of the viral proteins (72%) which can be associated with a host species, are found only in a single host species. Interestingly, some proteins can be found in up to 23 different host species. Those are capsid proteins and some replication associated proteins. 81.4% of viral TR-containing proteins have an eukaryotic host organism but only 3% have a bacterial host organism and 1% have archaeal hosts.

Proteins often have more than one single TR per protein. We couldn't find a relationship in figure 14 between the amount of TRs in viruses and their virus-host organisms. Because humans and their "virobiome" (43) are both great part of research and relatively many proteins are available for both of them, we show in figure 14 plot D the TR content of humans compared to human viruses. It can be seen, that overall viruses have more proteins without any TR than their host organisms. Virus-hosts have overall more proteins with only one single TR than viral proteins - which is not very distinct for Humans and their viruses. This could be due to the viruses having less proteins in general and belonging to less complex organisms, and, therefore, being less likely to include TRs.

DISCUSSION & CONCLUSION

With state of the art TR annotation methods, we found that about half of all known proteins have TRs. Eukaryotes and Viruses tend to have more TRs than Archae and Bacteria. This is in accordance with the findings of a previous census by Marcotte et al. (10).
We could further show, that the positive correlation between protein sequence length and the number of TR units which was previously observed (10) is still true, even with largley increased data.

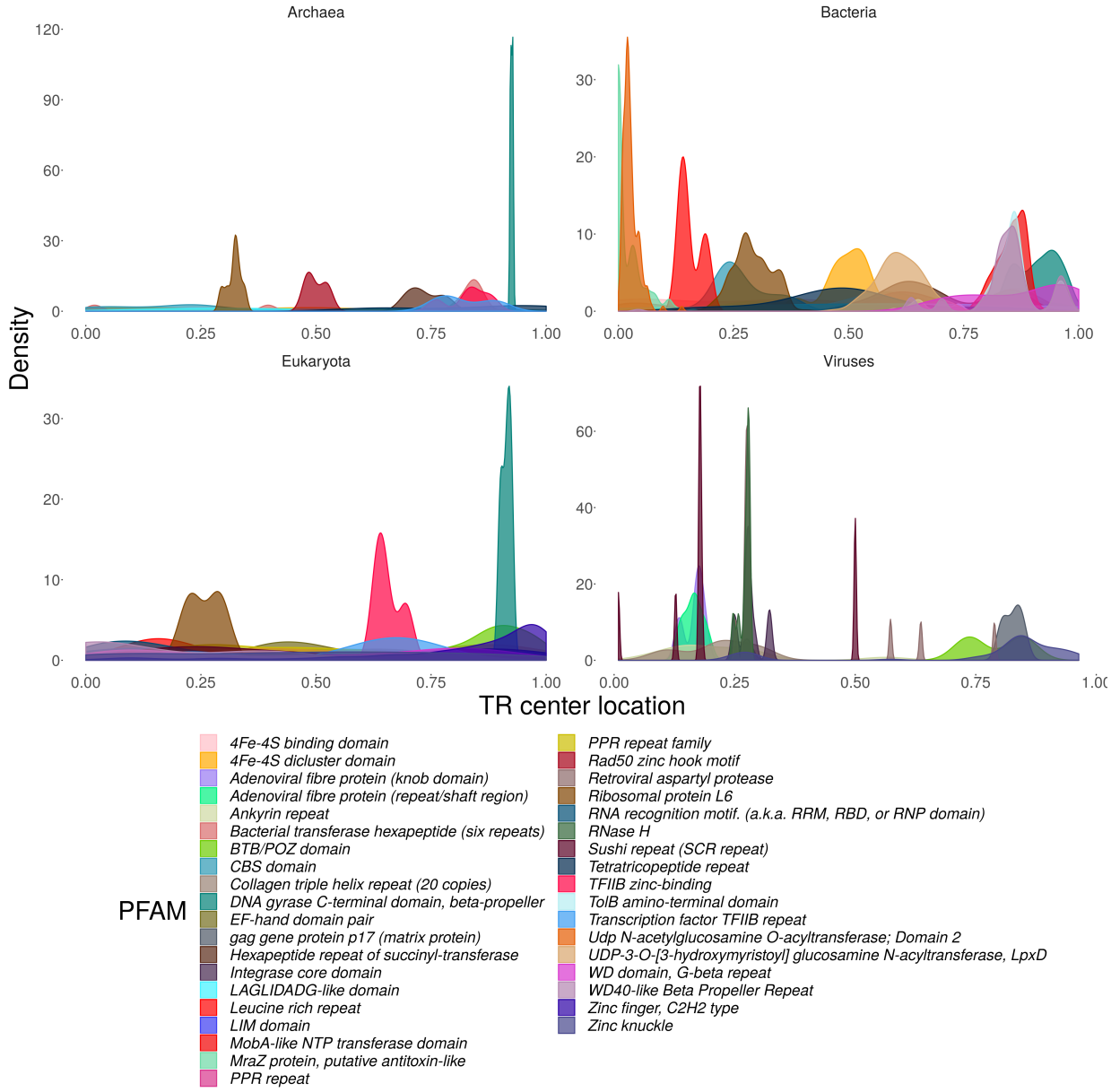


Figure 13. The ten PFAM with the most detected TRs for each Superkingdom are plotted according their normalized TR center location (see Methods) and number of site-specific TRs.

TRs Originatinate Through Duplication

TRs have significantly more amino acids which are associated with increased disorder propensity than protein sequences without TR. Protein domains with conserved disorder regions were gained during evolution trough alternative splicing methods, resulting in protein extension with existing exons which contain the highest degree of disorder regions. This suggests that exonization of previous noncoding regions could be an important mechanism for the addition of disordered segments of proteins (45, 46, 47, 48, 49). Conserved disorder regions evolve more rapidly than regions with defined structures and are known to show good properties in binding nucleic acids and in protein-protein interactions (50).

By duplication of such regions, these properties could be modulated resulting in the evolution of new TRs. Marcotte et al. reasoned that repeat expansion requires less energy than the initial repeat formation and that long repeats are preferentially duplicated. This supports our observation of an increased amount of TRs in proteins known for rapid expansion and diversification such as WD-40 domains and Ribosomal L6 protein family (51, 52).

Small TRs are most recurrent in all Superkingdoms. This might be because the small size seems to be a good trade-off of TR unit length and energy investment in duplication [TODO discuss this with Maria! Might be

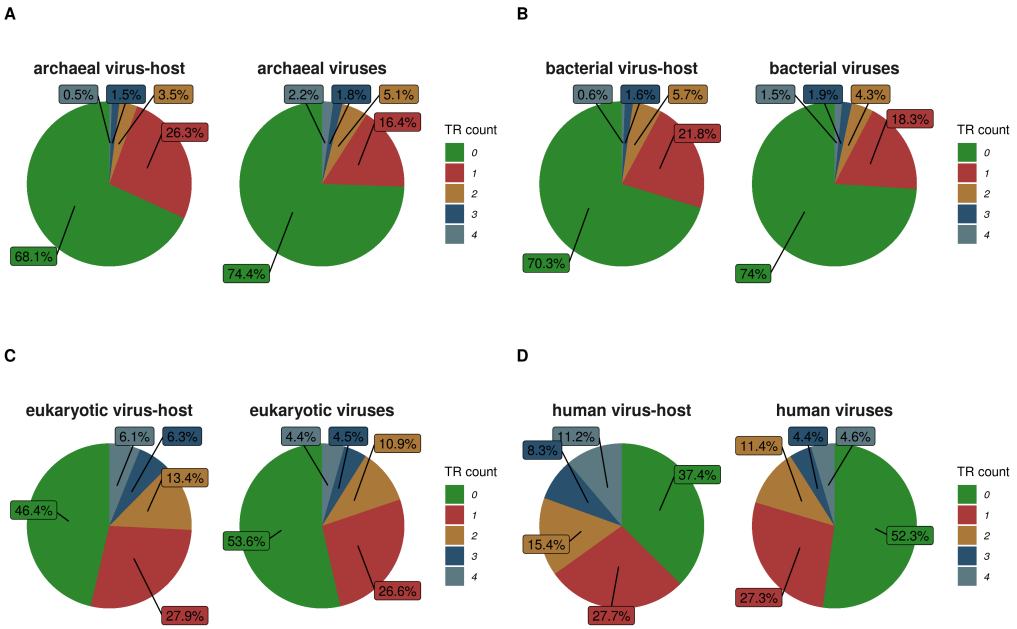


Figure 14. The ratio of the amount of TR per protein is shown as the number of TRs per protein divided by the total amount of proteins per group. As grouping factor we show in plots A, B, C the three superkingdoms each split into superkingdom specific viruses and their hosts. The analogous principle is shown in plot D but for human related viruses compared to the TR distribution in human proteins. It can be seen, that overall viruses have more proteins without any TR than their host organisms. Virus-hosts have overall more proteins with only one single TR - which is not very distinct for Humans and their viruses.

a thing for further investigation! (prove mathematically)] (36).

We found more homologous protein families between eukaryotes and prokaryotes (data not shown [would be in plotly graphs]) than between Eukaryotes and Viruses and vice versa. Proteins with chloroplastic or mitochondrial orgin, cluster toghether with prokaryotic proteins in regard of their mean protein length and TR content. Eukaryotic proteins with endosymbiotic origin tend to be shorter and with less TR. Proteins with origin in chloroplasts, are even shorter and with less TRs than mitochondrial proteins. In contrast to Marcottes (10) findings, we support the hypothesis of at least some proteins with TRs beeing involved in crucial mechanisms of prokaryotes which remained in endosymbiontes [TODO reformulate this sentence!]. We can see for example PPR repeats beeing overrepresented in eukaryotic proteins with endosymbiotic origin (mitochondria and chloroplast) and in prokaryotes. [TODO reformulate this sentence!].

TRs Are Involved in "Housekeeping" (44) Proteins

IDRs generally lack hydrophobic amino acids. The same holds for TRs [TODO refolumate this sentence!]. IDRs were therefore said of beeing unable to form well-organized hydrophobic cores that make up structured domains. However, TRs are known for both - beeing unstructured but also to fold into specific three dimensional shapes (16, 53). We therefore hyothesize, that if sequences of intrinsic disorder are repeatedly duplicated, they gain the ability to fold in specific tertiary structures [TODO: Is this possible? Discuss with Maria].

Eventhough, disorder content is highly species specific, it was shown, that Bacteria contain overall less IDR than Archaea and Viruses contain more IDR than both of them. For mitochondrial proteins no IDRs were found at all (54, 55, 56). That no IDR were found in mitochondrial proteins, supports our hypothesis, that certain, crucial TRs were generated before endosymbiosis of prokaryotes and persisted within the proteins. If initially disordered regions were accumulated in proteins through alternative splicing and by duplication established a stable tertiary structure, they might have been missed by the disorder detection methods [TODO check which kind of methods they used exactels (is was said "some" computational methods)!]

TRs were found to be enriched in proteins with functions in binding DNA and RNA (57). We could show that the TR location not only correlates with the location of the binding domains but also with the location of enriched intrinsic disorder. IDRs located on the N-terminus are common in DNA-binding proteins (58). C-terminal IDRs are associated with transcription factor repressor and activator activities. We could see that the TR locations in proteins from families involved in those mechanism corresponds to the findings of IDR. In general, TRs are located with a tendency to the N-terminal end of protein sequences. However, micro- and small TRs cluster to both termini. This finding might be explained by the

fact that domain TRs tend to be near the N-terminal end, hence they pull the location-distribution of TRs in a protein sequence to the N-terminus. More than half of the proteins with TRs, have more than one distinct TR and the TR-type can vary within the protein. Which might be due to different binding sites and/or different patterns on binding sites of proteins in hubs of protein-protein interaction networks

It was previously shown, that genomic TRs are involved in gene regulating mechanisms (34, 35). It would be interesting to understand how nucleic and proteomic TRs phyiscally interact with nucleic acids and in what way they differ.

Eukaryotic proteins tend to have more than one specific TR per protein. This goes hand in hand with the results that TR-regions are involved in gene regulation and signaling. It can thus be suggested that since more complex organisms require more genes to be regulated and more signaling happens, more TRs are found in eukaryotic organisms.

Observing the TR unit length and number, we detected some striking outliers belonging to host-colonialisation proteins of bacteria and viruses. Which indicate further, that the function of many proteins which are vital to survival is based on TRs.

The analysis of the overlap of tandemly repeated regions with regions marked as intrinsic disordered resulted in only a small overlap. But if they overlap, TRs fall mostly within the disordered regions. [TODO: think about this... why?].

Disorder Promoting Homorepeats Are Longer Than Order Promoting Homorepeats

We could show that homorepeats of amino acids with disorder promoting propensity generally have more repeat unit numbers in TR of IDR. Where polyH and polyN repeats, which are of low disorder promotion, appear unexpectedly frequent as long homo TRs. Therefore we could not only see that disorder promoting residues appear more frequently in disordered regions but also that their homo TRs are generally longer. These findings support the idea, that TRs play crucial roles in protein-protein interaction and signalling as well as the idea of folding in specific tertiary structures (i.e. amyloids).

Length variation in polyN and polyQ homo TRs play important roles in genetic diseases and are thought to be involved in the adaptation of organisms to their environments (62, 63, 64). It was shown that polyQ and polyN homo TRs rich proteins can be misfolded into amyloids which could be inferred with toxicity in yeast studies (65, 66, 67). There is evidence that those are caused through polyQ homo TRs stabilizing protein interactions (68). We found homo TRs more frequently in Eukaryotes than all other organisms. This findings are supported by different studies (68, 69). Schaefer et al. (68) suggest that proteins with long polyQ homo TRs are enriched in more complex species with a high amount of proteins involved in protein-interaction signaling and interact with proteins involved in transcriptional regulation. Which agrees with our generalized view of TRs beeing involved in transcriptional regulation and protein-protein interaction and with the observed increase amount of TRs in more complex organisms.

Excellent binding properties were attributed to polyH homo TRs (70) which can be finely regulated due to its physical and electrochemical properties such as the ability to change its charge, building charge gradients in β-sheets and has metal ion binding. PolyH homo TRs were found to be accumulated in the nuclear speckle involved in transcription processes. Given that TRs in general are involved in many transcriptional processes, an increased frequency of long polyH homo TRs can be explained through it's versatility and possibility for thight regulation of transcriptional processes in the cell cycle independet of its disorder propensity which can be fine-tuned in alteri through repeat expansion.

The Art and Design of TRs [TODO: put this as seperate conclusion section?]

CAVE: this are just thoughts of mine which I wanted to write down, but was not sure how, where and if at all they belong to a/this paper... But it might be interesting to summarize the already proposed methods for TR emergence including new ideas such as this:

It is known, that alternativly spliced exons are more likely to encode for intrinsically disordered proteins than for ordered.

If those regions, show evolutionary favourable, good binding affinity, those could be prone site of positive selection and/or are repeated to enhance its good binding affinities.

Upon the repetion of those sequences, structural changes occur. Those can be favorable or result after a certain length in unfavorable ones such as amyloids causing diseases (huntington).

-> this would explain, why in eukaryotes more TRs are found than in other Superkingdoms.

-> this would explain why TR are enriched in IDP and why they are found as part of IDR.

-> and why TRs are involved in RNA & DNA binding, transcription, signaling and protein-protein interactions (as those all can be positively selected if binding is enhanced). More complex organisms require thighter regulation of transcription and signaling.

However, the alternative splicing approach is restricted to Eukaryotes and thus is not the only TR-origination approach. There exist other ways leading to the emergence of TRs, which could also be based on genomic levels.

An Intimation of Overgeneralisation & for Critical Thinking

In our findings we could provide among new insights in the universe of proteins also support for previous results and hypotheses such as from Marcotte et al. (10). However, our study also warns against over-generalization of such observations. Clearly, protein sequences are highly heterogeneous in their

origin, content, structure and function across the diversity of organisms. Different biological processes may significantly contribute to TR origin, fixation and evolutionary mode. Therefore, we may observe exceptions from the general trend. For example, results and interpretations about TRs in Viruses and Archaea are based on a limited number of observations compared to eukaryotic and bacterial data which were and still are both under heavy investigation. Thus, more information is available for those and we therefore warn against an overgeneralisation because of this "investigation-bias" in the data.

MATERIAL & METHODS

Tandem repeat annotations

Amino acid tandem repeats (TRs) are neighboring sequence duplications in protein sequences. Depending on their repeat units, TRs vastly differ in their structural and biochemical properties: Homorepeats are repetitions of single amino acids (TR unit length $l=1$), we denote TRs with $l \leq 3$ as micro TRs, as they correspond to nucleic microsatellites. Further, we denote TRs with $4 \leq l < 15$ as small TRs, and TR with $l \geq 15$ as domain TRs.

STATISTICAL SIGNIFICANCE FILTER The shorter and the more diverged a TR, the harder it is to distinguish from a sequence without TR. To control the number of false-positive TR annotations in the dataset, we apply a model-based statistical significance filter (p=0.01), where the null hypothesis that the proposed TR units are evolutionary unrelated is tested against the alternative hypothesis that they are evolutionary related by duplication (4).

DE NOVO ANNOTATIONS All sequences were annotated with T-REKS (37), XSTREAM (38) and HHrepID (39) (default parameters). T-REKS and XSTREAM both excel at detecting short TRs, whilst HHrepID excels at detecting domain TRs.

TR ANNOTATIONS FROM PFAM DOMAINS PFAM domain annotation tags were retrieved from SwissProt. The corresponding sequence profile models were retrieved from PFAM (40), and converted to circular profile models, and used for tandem repeat annotation (5). A large number of annotated domains do not occur as TRs; these are filtered.

CONSENSUS ANNOTATIONS *de novo* annotations and PFAM annotations are subjected to a first filtering step ($p=0.1$, $n_{effective} > 1.9$). Next, for every sequence, the overlap of TR annotations is determined. To not filter small TRs within domain TRs, or TRs that overlap only in their flanks, overlap is not determined by shared amino acids. Instead, a strict version of the "shared ancestry" criterion is used: If two TR predictions share any two amino acids in the same column of their TR MSA, they are seen as the same TR. In this case, the *de novo* TR (in a tie with a PFAM TR) or the TR with lower p and higher divergence (in a tie between two *de novo* TRs) is removed.

To homogenize and refine all remaining *de novo* annotated TRs, they are converted to a circular profile hidden Markov model, reannotated (5), and subjected to stringent filtering (p=0.01).

We implemented the derived expressions in Python3 [The code is available]. The calculation is executed for every amino acid in all of SwissProt, and repeated for subsets of ordered and disordered regions.

Disorder

Intrinsically disordered regions often cause difficulties for experimental studies of protein structure, as these regions are inherently flexible, which can make proteins very difficult to crystallize, and hence X-ray diffraction analysis may be unfeasible. Even if X-ray crystals can be obtained or structure described via nuclear-magnetic resonance imaging (NMR), these data may still be hard to interpret due to random or missing values obtained for the disordered regions.

Based on what we know about intrinsic disorder: amino acid composition, hydropathy, capacity of polypeptides to form stabilizing contacts and other differences to known globular protein, - various computational methods have been developed to label each amino acid in a protein sequence as ordered or disordered.

While using these methods to study protein disorder and its evolution it is important to remember that they are limited to recognize patterns observed in experimentally annotated disorder and each predictor is tailored to identify a certain type of characteristics.

There is no standard definition of disorder and no large set of universally agreed disordered proteins. Moreover, different parts of proteins can be ordered or disordered under different conditions. It is therefore important to carefully annotate using different definitions of disorder.

DATA SOURCES Disorder annotations have been extracted from MobiDB covering 546,000 entries of UniProtKB/Swiss-Prot (Release 2014_07 (09. July 2014)). MobiDB provides consensus annotations as well as raw data from DisProt, PDB (missing residues in X-Ray and NMR) and 10 computational predictors.

PREDICTION METHODS Computational predictors assessed in our study include three ESpritz flavors, two IUPred flavors, two DisEMBL flavors, GlobPlot, VSL2b and JRONN. Computational methods analizing protein sequence usually provide a per-residue probability scoring of protein disorder, with a cutoff of 0.5 to be considered disordered.

MACHINE LEARNING The following methods are based on machine learning and trained on various experimentally obtained data: ESpritz ensemble of disorder predictors is based on bidirectional recursive neural networks and trained on three different flavors of disorder: Disprot, Xray and NMR flexibility.

DisEMBL-465 , DisEMBL-HL predictors are focusing on shorter disordered regions, - loops with high B-factor

(high flexibility), defining disorder as "hot loops", i.e., coils with high temperature factors.

JRONN is a regional order neural network (RONN) software that employs a bio-basis sequence similarity function that was initially developed for prediction of protease cleavage sites.

VSL2b predictor addresses the differences in disordered regions of different length, modelling short and long disordered regions separately and is using a linear SVM approach for predictions.

Data Analysis

The data analysis pipeline was performed in R version 3.6.0 (2019-04-26) (61) [code available online].

TR LOCATION NORMALIZATION If the TR covers a significant fraction of the protein, then it's center necessarily falls near the middle. To avoid a center-bias (espescially for domain TRs), coming from boundary effects from the simple center/length metric, we can compensate for this by normalizing over only the valid center locations:

$$x = \frac{center - \frac{l_{eff} \cdot n_{eff}}{2}}{N - l_{eff} \cdot n_{eff}} \quad (\mathbf{1})$$

With N representing the number of amino acids of a protein (protein length). The TR size in terms of number of amino acids is calculated by the number of repeat units n_{eff} and the repeat unit length l_{eff}. *center* is the position of the AA in the middel of the TR of size $l_{eff} \cdot n_{eff}$. We further filtered for entries with the main denominator > 0.

HOMOREPEAT ANNOTATIONS To compare expected and empirical number of homorepeats in SwissProt, we exactly counted the number of runs of all lengths and all amino acids in SwissProt. We repeated this exact count for bounded subsets of SwissProts, such as disordered and ordered regions, according to different definitions of either.

EXPECTED NUMBER OF HOMOREPEATS We want to derive the expected number of homorepeats of amino acid a with n repeat units in a random sequence of length s, given the amino acid frequency $p(a)$. Mathematically, this problem corresponds to sequential runs of successes in a Bernoulli trial. The probability of amino acid a equates to the probability of a success, and the expected values and variances can be derived for all sequences or subsequences of different lengths in the sequence set. Exact solutions to the expected value and variance of the number of runs of a given length in a bounded sequence of length are derived in, e.g., (33).

CORRELATION ANALYSIS OF MEAN PROTEIN LENGTH AND TR-FRACTION A one-sided Pearson product-momentum correlation coefficient test was computed to assess the relationship between the protein length and the fraction of proteins with TRs for each type of TR seperately with 12 degrees of freedom. The null-hypothesis that there is no correlation $H_0 : \rho_0 = 0$ was tested against the alternative hypothesis $H_A : \rho_A > 0$ for observing a positive correlation on an significance level of $\alpha = 0.001$.

CORRELATION ANALYSIS OF AMINO ACID ABUNDANCY AND IT'S DISORDER PROPENSITY A two-sided Spearman's rank correlation analysis was performed on a significance level of $\alpha = 0.05$ to assess the relationship between the amino acid disorder propensity (disorder propensity values from (60)) and seperately for the amino acid frequency in TRs, for all protein sequences and for all protein sequences without the ones, which show at least one TR.

ACKNOWLEDGEMENTS

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.

Conflict of interest statement. None declared.

REFERENCES

1. Van Belkum, A., Scherer, S., Van Alphen, L., Verbrugh, H. (1998) Short-Sequence DNA Repeats in Prokaryotic Genomes. *Microbiology and Molecular Biology Reviews*, **62**, 275-293.
2. Richard, G., Kerrest, A. and Dujon, B. (2008) Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. *Microbiology and Molecular Biology Reviews*, **72**, 686-727.
3. Lim, Kian Guan et al. (2013). Review of tandem repeat search tools: A systematic approach to evaluating algorithmic performance. *Briefings in Bioinformatics*, **14**, 67–81.
4. Schaper, E., Kajava, A., Hauser, A. and Anisimova, M. (2012) Repeat or not repeat? Statistical validation of tandem repeat prediction in genomic sequences. *Nucleic Acids Research*, **40**, 10005-10017.
5. Schaper, E., Gascuel, O. and Anisimova, M. (2014) Deep Conservation of Human Protein Tandem Repeats within the Eukaryotes. *Molecular Biology and Evolution*, **31**, 1132-1148.
6. Ellegren, Hans (2004). Microsatellites: Simple sequences with complex evolution. *Nature Reviews Genetics*, **5**, 435–445.
7. Nithianantharajah, Jess and Anthony J. Hannan (2007). Dynamic mutations as digital genetic modulators of brain development, function and dysfunction. *BioEssays*, **29**, 525–535.
8. Javadi, Y. and Itzhaki, L. (2013) Tandem-repeat proteins: regularity plus modularity equals design-ability. *Current Opinion in Structural Biology*, **23**, 622-631.
9. Di Domenico, T., Potenza, E., Walsh, I., Gonzalo Parra, R., Giollo, M., Minervini, G., Piovesan, D., Ihsan, A., Ferrari, C. and Kajava, A. et al. (2013) RepeatsDB: a database of tandem repeat protein structures. *Nucleic Acids Research*, **42**, D352-D357.
10. Marcotte, E., Pellegrini, M., Yeates, T. and Eisenberg, D. (1999) A census of protein repeats. *Journal of Molecular Biology*, **293**, 151-160.
11. Anisimova, M., Pečerska, J. and Schaper, E. (2015) Statistical Approaches to Detecting and Analyzing Tandem Repeats in Genomic Sequences. *Frontiers in Bioengineering and Biotechnology*, **3**.
12. Schaper, E., Korsunsky, A., Pečerska, J., Messina, A., Murri, R., Stockinger, H., Zoller, S., Xenarios, I. and Anisimova, M. (2015) TRAL: tandem repeat annotation library. *Bioinformatics*, **31**, 3051-3053.
13. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A., Poux, S., Bougueleret, L. and Xenarios, I. (2016) UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Plant Bioinformatics*, **1374**, 23-54.
14. Szalkowski, A. and Anisimova, M. (2013) Graph-based modeling of tandem repeats improves global multiple sequence alignment. *Nucleic Acids Research*, **41**, e162-e162.
15. Ekman, D., Light, S., Björklund, Å. and Elofsson, A. (2006) *Genome Biology*, **7**, R45.
16. Kajava, A. (2012) Tandem repeats in proteins: From sequence to structure. *Journal of Structural Biology*, **179**, 279-288.
17. Jorda, J., Xue, B., Uversky, V. and Kajava, A. (2010) Protein tandem repeats - the more perfect, the less structured. *FEBS Journal*, **277**, 2673-2682.
18. Tompa, P. (2003) Intrinsically unstructured proteins evolve by repeat expansion. *BioEssays*, **25**, 847-855.
19. Simon, M. and Hancock, J. (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biology*, **10**, R59.
20. Light, S., Sagit, R., Sachenkova, O., Ekman, D. and Elofsson, A. (2013) Protein Expansion Is Primarily due to Indels in Intrinsically Disordered Regions. *Molecular Biology and Evolution*, **30**, 2645-2653.
21. Li, C., Ng, M., Zhu, Y., Ho, B. and Ding, J. (2003) Tandem repeats of Sushi3 peptide with enhanced LPS-binding and -neutralizing activities. *Protein Engineering Design and Selection*, **16**, 629-635.
22. Usdin, K. (2008) The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. *Genome Research*, **18**, 1011-1019.
23. Madsen, B., Villesen, P. and Wiuf, C. (2008) Short Tandem Repeats in Human Exons: A Target for Disease Mutations. *BMC Genomics*, **9**, 410.
24. Dashnow, H., Lek, M., Phipson, B., Halman, A., Sadedin, S., Lonsdale, A., Davis, M., Lamont, P., Clayton, J. and Laing, N. et al. (2018) STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biology*, **19**, 121.
25. Fertin, G., Jean, G., Radulescu, A. and Rusu, I. (2015) Hybrid de novo tandem repeat detection using short and long reads. *BMC Medical Genomics*, **8**, S5.
26. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **45**, D158-D169.
27. Rollins, R. (2005) Large-scale structure of genomic methylation patterns. *Genome Research*, **16**, 157-163.
28. Hannan, A. (2018) Tandem repeats mediating genetic plasticity in health and disease. *Nature Reviews Genetics*, **19**, 286-298.
29. Nelson, K., Fleischmann, R., DeBoy, R., Paulsen, I., Fouts, D., & Eisen, J. et al. (2003) Complete Genome Sequence of the Oral Pathogenic Bacterium Porphyromonas gingivalis Strain W83. *Journal Of Bacteriology*, **185(18)**, 5591-5601.
30. Han, N., Whitlock, J., & Progulske-Fox, A. (1996). The hemagglutinin gene A (hagA) of Porphyromonas gingivalis 381 contains four large, contiguous, direct repeats. *Infection and immunity*, **64(10)**, 4000-7.
31. Harada, N., Iijima, S., Kobayashi, K., Yoshida, T., Brown, W., & Hibi, T. et al. (1997) Human IgGFc Binding Protein (FcBP) in Colonic Epithelial Cells Exhibits Mucin-like Structure. *Journal Of Biological Chemistry*, **272(24)**, 15232-15241.
32. Dunker, A., Lawson, J., Brown, C., Williams, R., Romero, P., & Oh, J. et al. (2001) Intrinsically disordered protein. *Journal Of Molecular Graphics And Modelling*, **19(1)**, 26-59.
33. Makri, F. S., & Psillakis, Z. M. (2011) On success runs of a fixed length in Bernoulli sequences: Exact and asymptotic results. *Computers & Mathematics with Applications*, **61(4)**, 761–772.
34. Bilgin Sonay, T., Koletou, M., & Wagner, A. (2015) A survey of tandem repeat instabilities and associated gene expression changes in 35 colorectal cancers. *BMC Genomics*, **16(1)**.
35. Theriot, J. (2013) Why are bacteria different from eukaryotes?. *BMC Biology*, **11(1)**.
36. Andrade, M., Perez-Iratxeta, C., & Ponting, C. (2001) Protein Repeats: Structures, Functions, and Evolution. *Journal Of Structural Biology*, **134(2-3)**, 117-131.
37. Jorda, J., & Kajava, A. (2009) T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics*, **25(20)**, 2632-2638.
38. Newman, A., & Cooper, J. (2007) XSTREAM: A practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics*, **8(1)**.
39. Biegert, A., & Söding, J. (2008) De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*, **24(6)**, 807-814.
40. Finn, R., Coggill, P., Eberhardt, R., Eddy, S., Mistry, J., & Mitchell, A. et al. (2015) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, **44(D1)**, D279-D285.
41. Walsh, D., Mathews, M., & Mohr, I. (2012) Tinkering with Translation: Protein Synthesis in Virus-Infected Cells. *Cold Spring Harbor Perspectives In Biology*, **5(1)**, a012351-a012351.
42. Stern-Ginossar, N., Thompson, S., Mathews, M., & Mohr, I. (2018) Translational Control in Virus-Infected Cells. *Cold Spring Harbor Perspectives In Biology*, **11(3)**, a033001.
43. Ruff, M. (2017) Virobiome Derived Peptide T: Anti-Inflammatory Peptides for Treating Neuro-Aids and Neurodegenerative Diseases. *Journal Of Microbiology & Experimentation*, **5(2)**.
44. Ferguson, R., Carroll, H., Harris, A., Maher, E., Selby, P., & Banks, R. (2005) Housekeeping proteins: A preliminary study illustrating some limitations as useful references in protein expression studies. *PROTEOMICS*, **5(2)**, 566-571.
45. Kriventseva, E., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M., & Sunyaev, S. (2003) Increase of functional diversity by alternative splicing. *Trends In Genetics*, **19(3)**, 124-128.
46. Romero, P., Zaidi, S., Fang, Y., Uversky, V., Radivojac, P., & Oldfield, C. et al. (2006) Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in

multicellular organisms. *Proceedings Of The National Academy Of Sciences*, **103(22)**, 8390-8395.
47. Hegyi, H., Kalmar, L., Horvath, T., & Tompa, P. (2010) Verification of alternative splicing variants based on domain integrity, truncation length and intrinsic protein disorder. *Nucleic Acids Research*, **39(4)**, 1208-1219.
48. Buljan, M., Chalancon, G., Eustermann, S., Wagner, G., Fuxreiter, M., Bateman, A., & Babu, M. (2012) Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Molecular Cell*, **46(6)**, 871-883.
49. Barbosa-Morais, N., Irimia, M., Pan, Q., Xiong, H., Gueroussov, S., & Lee, L. et al. (2012) The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science*, **338(6114)**, 1587-1593.
50. Chen, J., Romero, P., Uversky, V., & Dunker, A. (2006) Conservation of Intrinsic Disorder in Protein Domains and Families: II. Functions of Conserved Disorder. *Journal Of Proteome Research*, **5(4)**, 888-898.
51. Smith, T., Gaitatzes, C., Saxena, K., & Neer, E. (1999) The WD repeat: a common architecture for diverse functions. *Trends In Biochemical Sciences*, **24(5)**, 181-185.
52. Golden, B., Ramakrishnan, V., & White, S. (1993) Ribosomal protein L6: structural evidence of gene duplication from a primitive RNA binding protein. *The EMBO Journal*, **12(13)**, 4901-4908.
53. Paladin, L., Hirsh, L., Piovesan, D., Andrade-Navarro, M., Kajava, A., & Tosatto, S. (2016) RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures. *Nucleic Acids Research*, **45(D1)**, D308-D312.
54. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R., Daughdrill, G., & Dunker, A. et al. (2014) Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews*, **114(13)**, 6589-6631.
55. Pavlović -Lažetić, G., Mitić, N., Kovačević, J., Obradović, Z., Malkov, S., & Beljanski, M. (2011) Bioinformatics analysis of disordered proteins in prokaryotes. *BMC Bioinformatics*, **12(1)**, 66.
56. Pentony, M., & Jones, D. (2009) Modularity of intrinsic disorder in the human proteome. *Proteins: Structure, Function, And Bioinformatics*, **78(1)**, 212-221.
57. Lobley, A., Swindells, M., Orengo, C., & Jones, D. (2007) Inferring Function Using Patterns of Native Disorder in Proteins. *Plos Computational Biology*, **3(8)**, e162.
58. Vuzman, D., Azia, A., & Levy, Y. (2010) Searching DNA via a "Monkey Bar" Mechanism: The Significance of Disordered Tails. *Journal Of Molecular Biology*, **396(3)**, 674-684.
59. Light, S., Basile, W., & Elofsson, A. (2014) Orphans and new gene origination, a structural and evolutionary perspective. *Current Opinion In Structural Biology*, **26**, 73-83.
60. Uversky, V. (2013) The alphabet of intrinsic disorder. *Intrinsically Disordered Proteins*, **1(1)**, e24684.
61. R Development Core Team (2019) R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria.
62. Caprioli, M., Ambrosini, R., Boncoraglio, G., Gatti, E., Romano, A., & Romano, M. et al. (2012) Clock Gene Variation Is Associated with Breeding Phenology and Maybe under Directional Selection in the Migratory Barn Swallow. *Plos ONE*, **7(4)**, e35140.
63. Undurraga, S., Press, M., Legendre, M., Bujdoso, N., Bale, J., & Wang, H. et al. (2012) Background-dependent effects of polyglutamine variation in the Arabidopsis thaliana gene ELF3. *Proceedings Of The National Academy Of Sciences*, **109(47)**, 19363-19367.
64. Michael, T., Park, S., Kim, T., Booth, J., Byer, A., & Sun, Q. et al. (2007) Simple Sequence Repeats Provide a Substrate for Phenotypic Variation in the Neurospora crassa Circadian Clock. *Plos ONE*, **2(8)**, e795.
65. Kochneva-Pervukhova, N., Alexandrov, A., & Ter-Avanesyan, M. (2012) Amyloid-Mediated Sequestration of Essential Proteins Contributes to Mutant Huntingtin Toxicity in Yeast. *Plos ONE*, **7(1)**, e29832.
66. Alexandrov, A., & Ter-Avanesyan, M. (2013) Could yeast prion domains originate from polyQ/N tracts? *Prion*, *7(3)*, 209-214.
67. Gong, H., Romanova, N., Allen, K., Chandramowlishwaran, P., Gokhale, K., & Newnam, G. et al. (2012) Polyglutamine Toxicity Is Controlled by Prion Composition and Gene Dosage in Yeast. *Plos Genetics*, **8(4)**, e1002634.
68. Schaefer, M., Wanker, E., & Andrade-Navarro, M. (2012) Evolution and function of CAG/polyglutamine repeats in protein–protein interaction networks. *Nucleic Acids Research*, **40(10)**, 4273-4287.
69. Huntley, M., & Golding, G. (2002) Simple sequences are rare in the Protein Data Bank. *Proteins: Structure, Function, And Genetics*, **48(1)**, 134-140.
70. Salichs, E., Ledda, A., Mularoni, L., Albá, M., & de la Luna, S. (2009) Genome-Wide Analysis of Histidine Repeats Reveals Their Role in the Localization of Human Proteins to the Nuclear Speckles Compartment. *Plos Genetics*, **5(3)**, e1000397.

Supplementary Materials: A new census of protein tandem repeats: fun with disorder.

INTRODUCTION

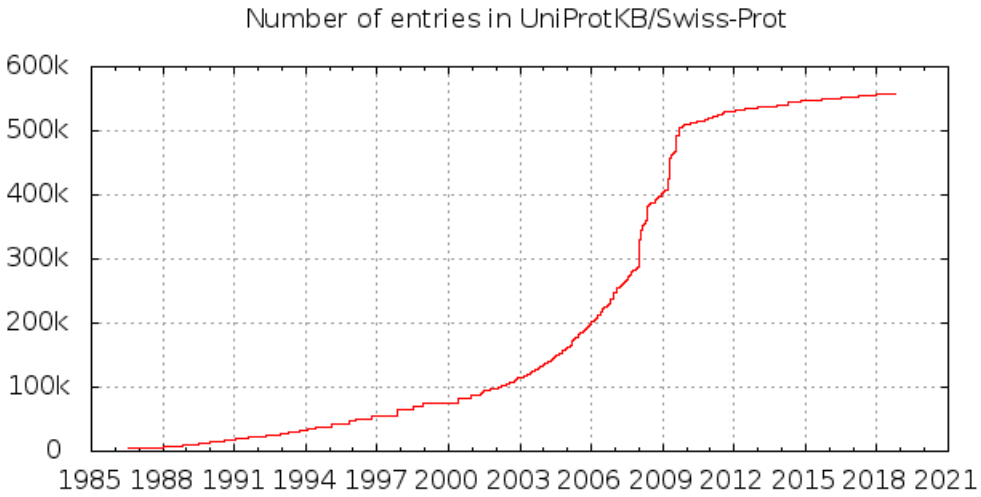


Figure S1. Summary of the growth of UniProtKB/Swiss-Prot protein knowledgebase. The last protein census dates back to the year 1999 (10). Since then, the entries in the UniProtKB/Swiss-Prot protein knowledgebase are grown more than seven fold. Figure from release 2018_09 statistics https://web.expasy.org/docs/relnotes/relstat.html, retrieved 2018/10/17.

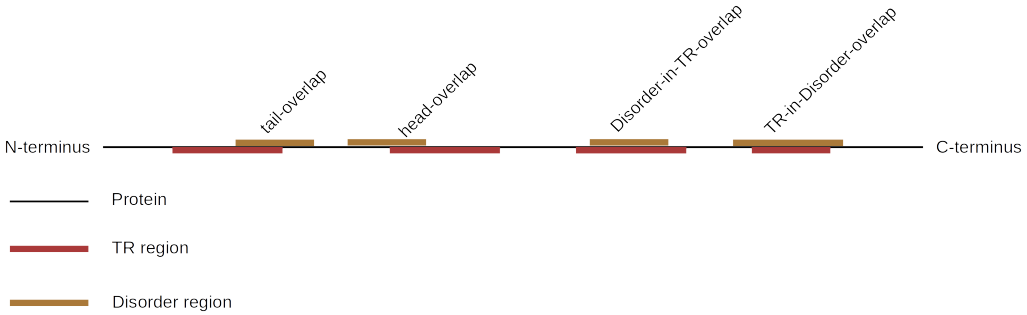
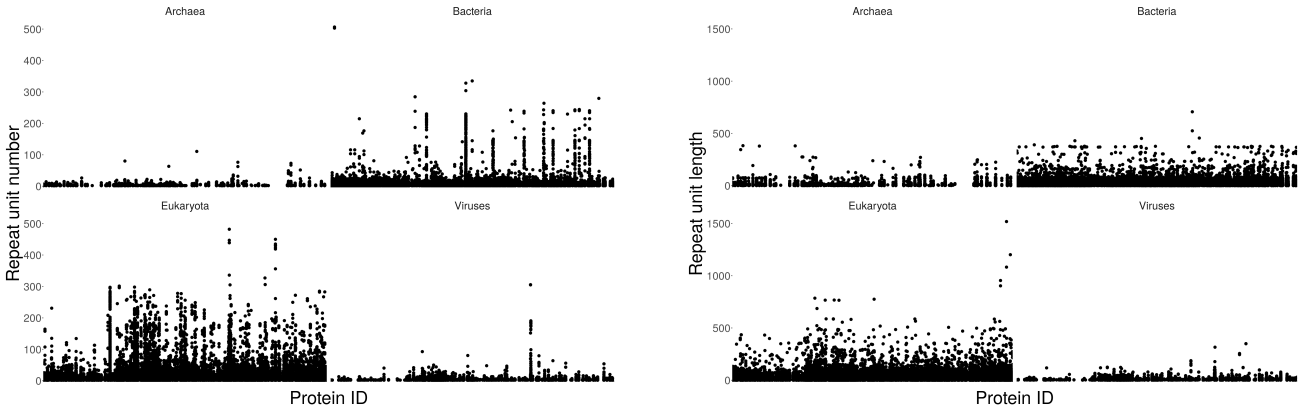


Figure S2. Overlap regions in proteins with intinsic disorder and tandem repeats. We distinguish four different overlaps of IDR with TRs: *tail-overlap* where IDR begin within the TR-sequence and finishes after the TR-region. In contrast, we call *head-overlaps* overlap regions when the IDR begins before the TR-sequence and finishes within. If the IDR lies within a TR sequence, we call it *Disorder-in-TR* and *TR-in-Disorder-overlap* if the TR-region lies within the IDR.

RESULTS



(a) Distribution of TRs by TR size.

(b) Distribution of TRs by TR unit length

Figure S3. Distribution of TRs by the number of repetition of the minimal TR unit (A) and their unit length (B). Showing that Bacteria and Eukaryota tend to have more repetitions and longer TR units than Archaea and Viruses. Where eukaryotic proteins thend to be more uniformely distributed than TRs from baceria. One can see in (b) that Eukaryota have certain proteins with specially long TR units.

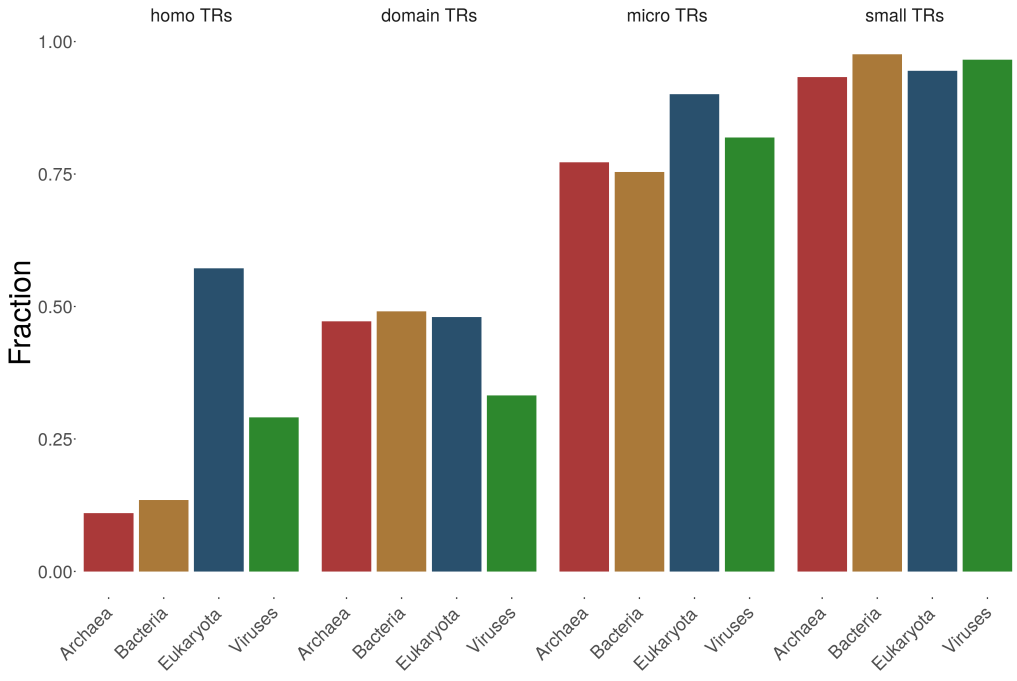


Figure S4. Proteins with ≥ 4 distinct TR regions are sorted by their TR type and shown kingdomwise. One can clearly see, that over all kingdoms small TRs dominate in proteins with many distinct regions.

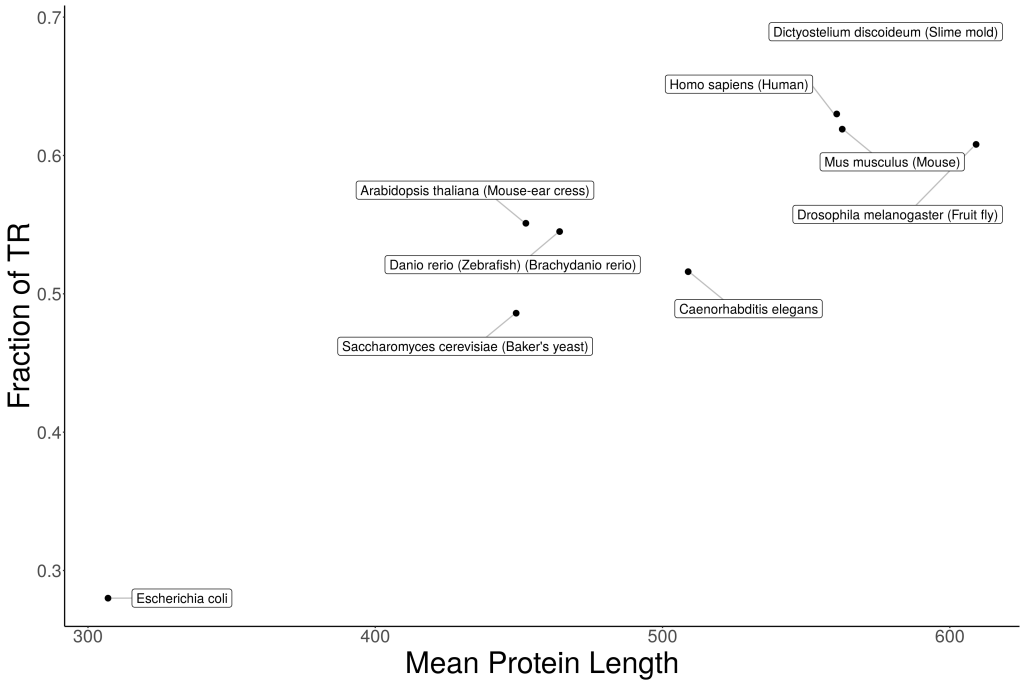


Figure S5. The fraction of proteins containing TRs over all protein entries in UniProtKB/Swiss-Prot is shown for a selection of species and displayed as function of the mean protein length.

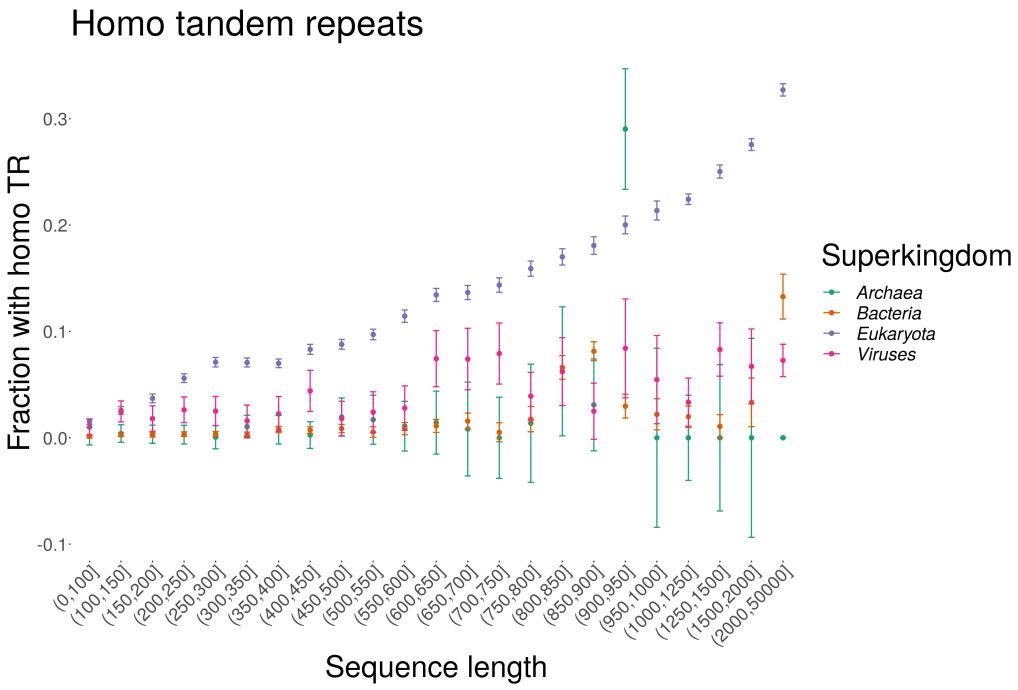


Figure S6. The fraction of proteins with homo TRs as a function of sequence length by kingdom resulting in a linear relationship.

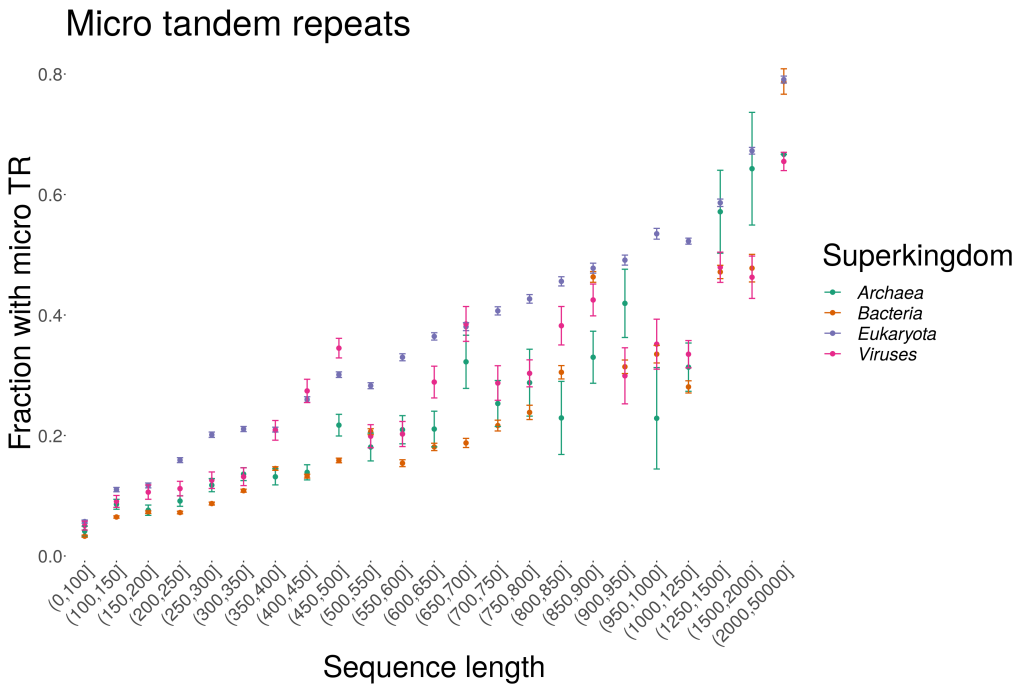


Figure S7. The fraction of proteins with micro TRs as a function of sequence length by kingdom resulting in a linear relationship.

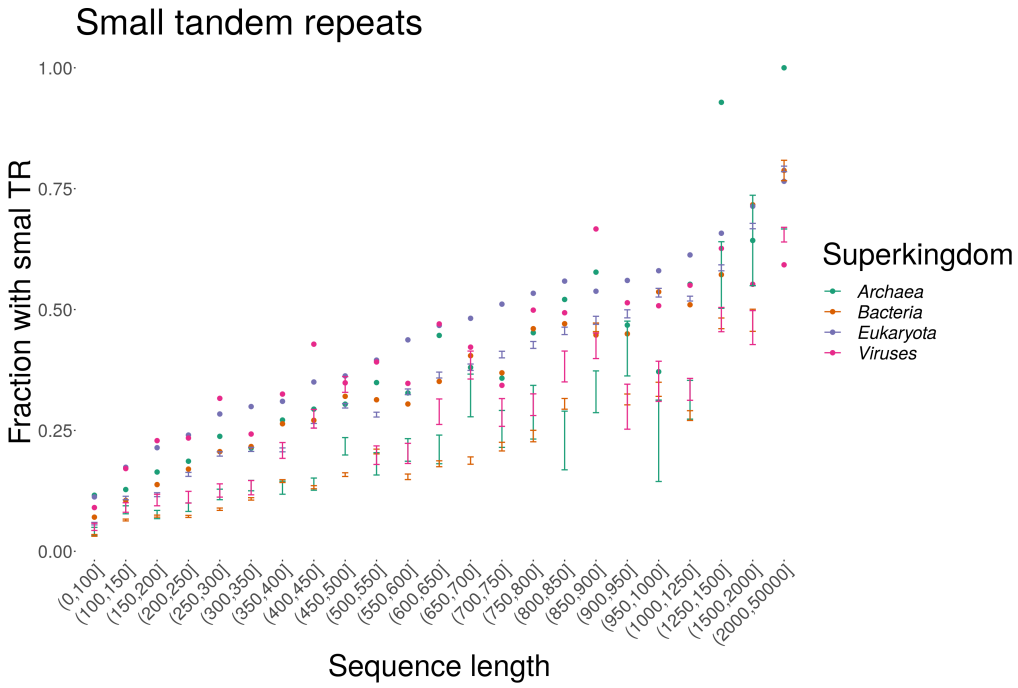


Figure S8. The fraction of proteins with small TRs as a function of sequence length by kingdom resulting in a linear relationship.

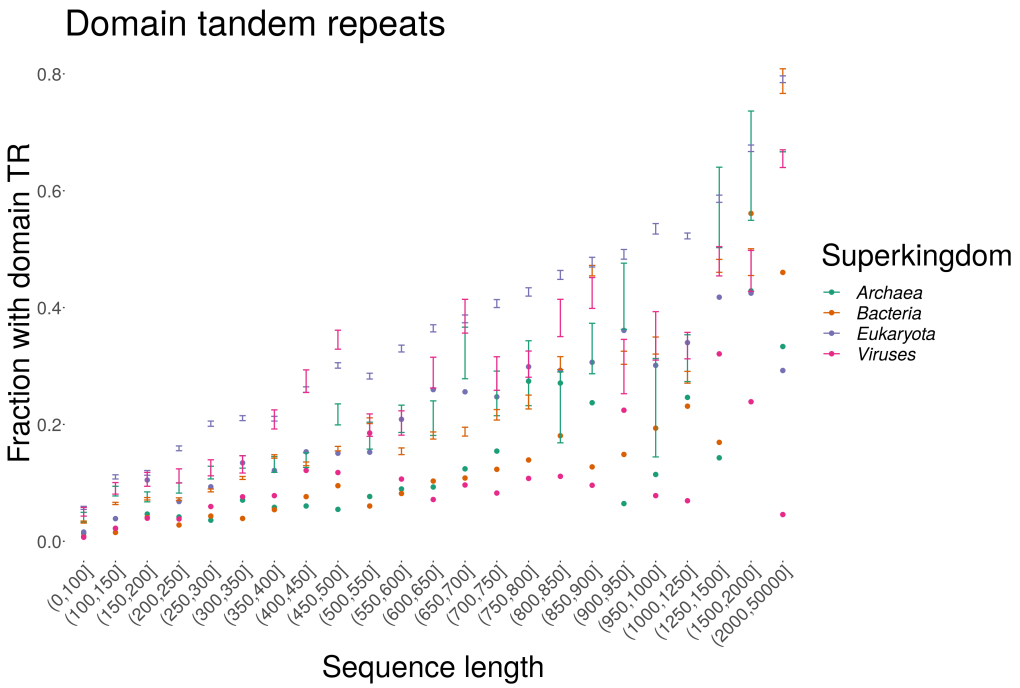


Figure S9. The fraction of proteins with domain TRs as a function of sequence length by kingdom resulting in a linear relationship.

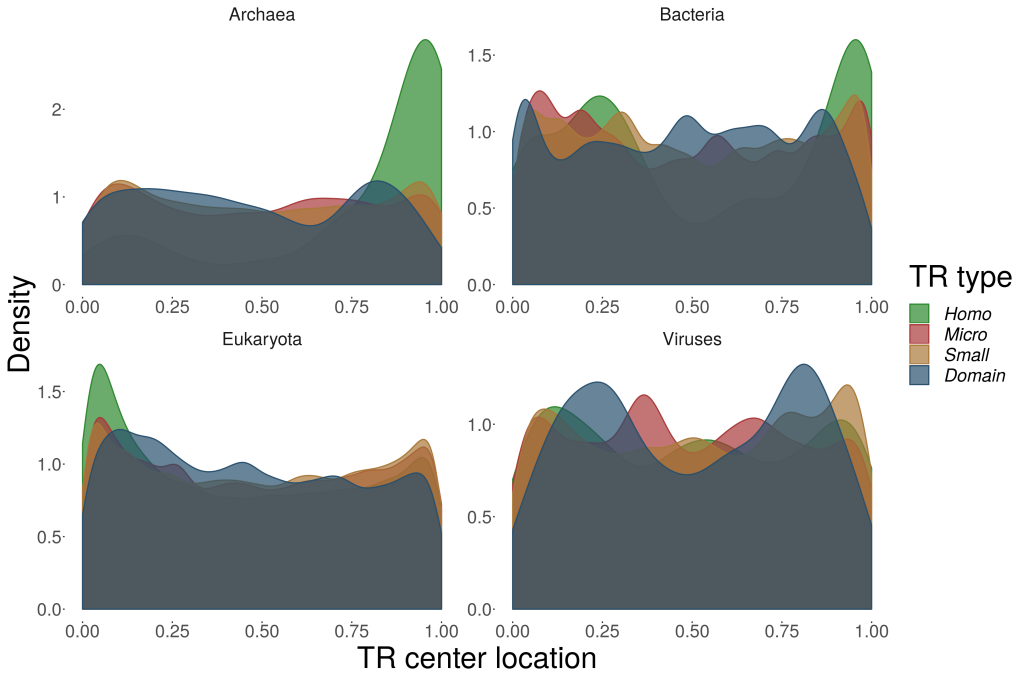


Figure S10. Density plots for the relative positions of TRs within proteins for four Superkindoms. The relative position referes with 0 to the N-terminus and with 1 to the C-terminus of a protein. Colours indicate repeat unit lengths. Interestingly, short TRs are biased towards the flanks of the protein. In particular for Eukaryotes, there is a clear correlation between TR unit length and location bias to the protein flanks. For Eukaryotes, tandem repeats are particularly prevalent in the N-terminal protein flank. Homorepeats in Archaea and, to a lesser degree, in Bacteria show a strong bias to the C-terminal protein flank.

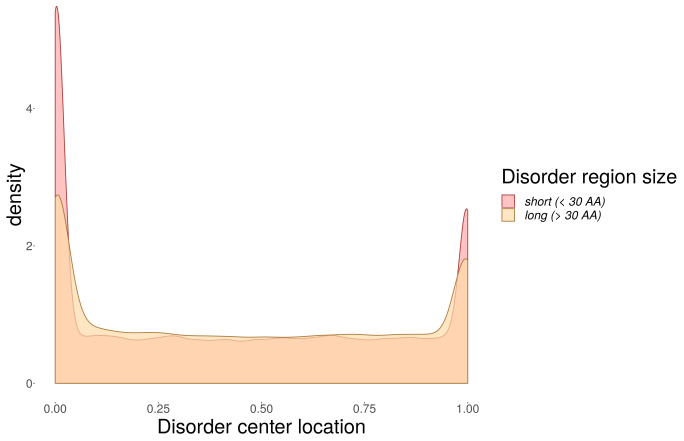


Figure S11. Density plots of position of disorder regions within the protein for four Superkingdoms. Both short and long disorder regions tend to cluster towards the flank of the protein, to the N-terminal specifically, with the trend being somewhat weaker in Eukaryotes.

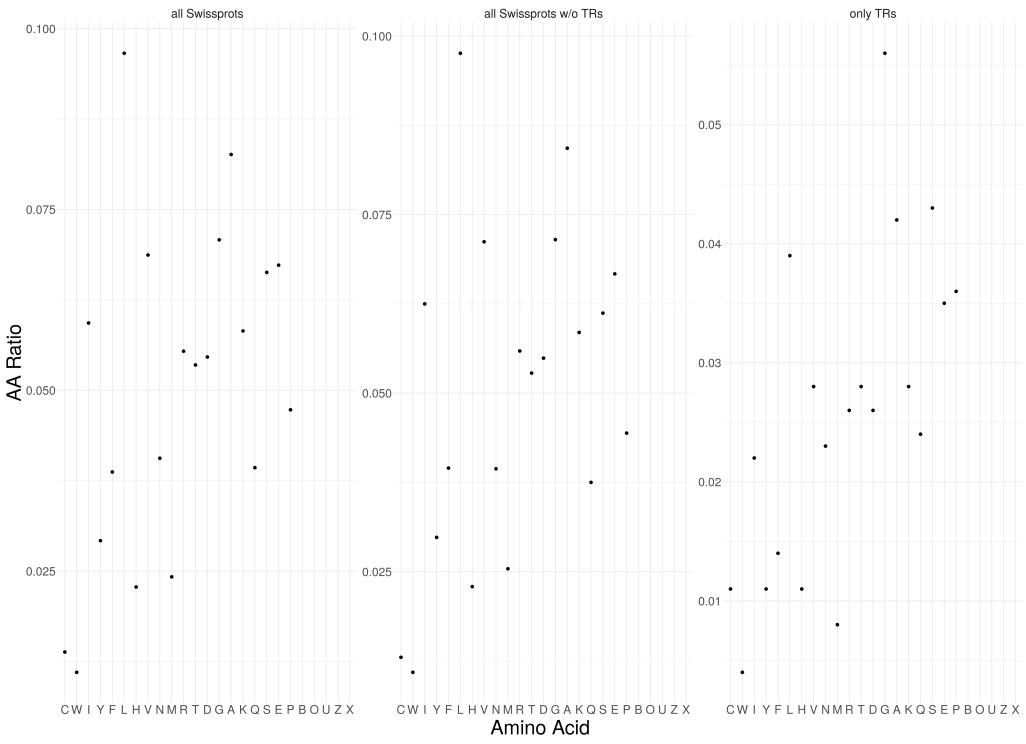


Figure S12. The amino acid ratio was calculated by the number of appearance of each amino acid divided by the overall number of amino acids per category and plotted against the amino acids in increasing disorder promoting potential. The group of all SwissProts represents all protein sequences from SwissProt. Of those, all proteins which have at least one detected TR were subtracted resulting in the group 'all Siwssprots w/o TRs'. For the group 'only TRs' was calculated by the multiple sequence alignment of the TRs. For the amino acids B, O, U, Z and X was no disorder potential available. One can see that the amino acid ratio of TR sequences shows a positive linear relationship with increasing disorder propensity. Disorder promoting residues seem to appear more often in TR sequences compared to overall protein sequences and to proteins without TRs.

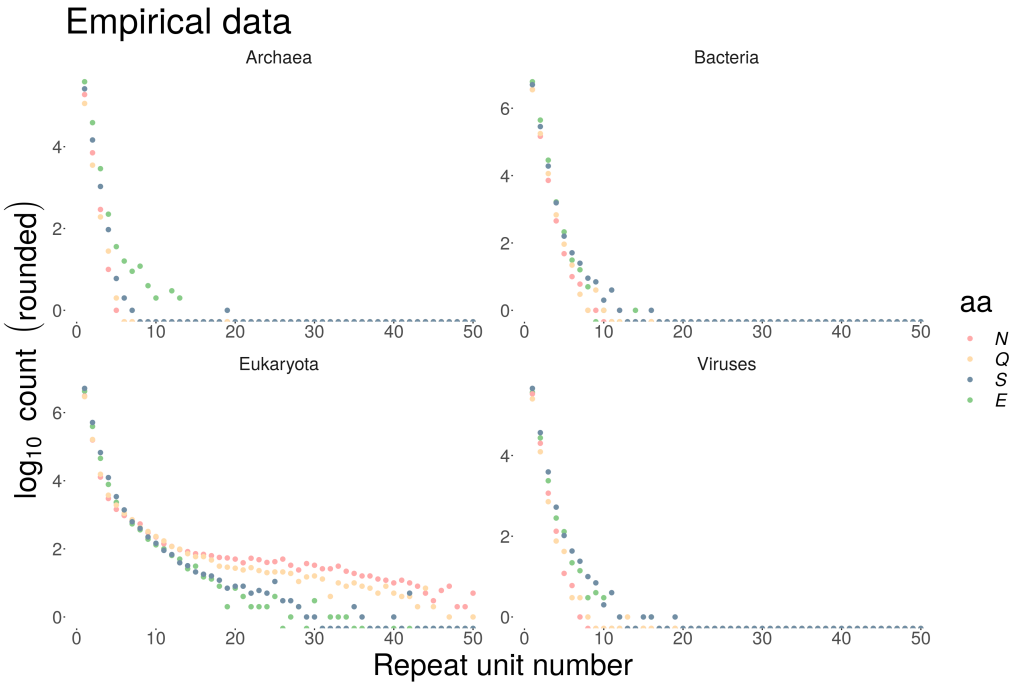


Figure S13. Count of homorepeats in Swiss-Prot in four Superkingdoms for different repeat unit number ($n <= 50$, equivalent to repeat length) for amino acids E, S, N and Q. Homorepeats with large n seem to mostly pertain to the Eukaryotes.

PFAM Name	PFAM Desc	PFAM Acc	count
Archaea			
TFIIB	Transcription factor TFIIB repeat	PF00382	35
CBS	CBS domain	PF00571	22
Fer4	4Fe-4S binding domain	PF00037	16
Fer4_7	4Fe-4S dicluster domain	PF12838	13
LAGLIDADG_3	LAGLIDADG-like domain	PF14528	11
Hexapep	Bacterial transferase hexapeptide (six repeats)	PF00132	9
TF_Zn_Ribbon	TFIIB zinc-binding	PF08271	9
Ribosomal_L6	Ribosomal protein L6	PF00347	7
Rad50_zn_hook	Rad50 zinc hook motif	PF04423	7
Fer4_10	4Fe-4S dicluster domain	PF13237	7
Bacteria			
Hexapep	Bacterial transferase hexapeptide (six repeats)	PF00132	928
MraZ	MraZ protein, putative antitoxin-like	PF02381	320
Ribosomal_L6	Ribosomal protein L6	PF00347	317
NTP_transf_3	MobA-like NTP transferase domain	PF12804	244
Hexapep_2	Hexapeptide repeat of succinyl-transferase	PF14602	223
PD40	WD40-like Beta Propeller Repeat	PF07676	164
Acetyltransf_11	Udp N-acetylglucosamine O-acyltransferase; Domain 2	PF13720	158
LpxD	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase, LpxD	PF04613	127
TolB_N	TolB amino-terminal domain	PF04052	115
DNA_gyraseA_C	DNA gyrase C-terminal domain, beta-propeller	PF03989	100
Eukaryota			
WD40	WD domain, G-beta repeat	PF00400	1449
zf-C2H2	Zinc finger, C2H2 type	PF00096	828
LRR_8	Leucine rich repeat	PF13855	587
EF-hand_7	EF-hand domain pair	PF13499	520
RRM_1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	PF00076	413
LIM	LIM domain	PF00412	260
PPR	PPR repeat	PF01535	226
PPR_2	PPR repeat family	PF13041	225
TPR_1	Tetratricopeptide repeat	PF00515	184
Collagen	Collagen triple helix repeat (20 copies)	PF01391	181
Viruses			
zf-CCHC	Zinc knuckle	PF00098	56
Gag_p17	gag gene protein p17 (matrix protein)	PF00540	37
RVP	Retroviral aspartyl protease	PF00077	13
Ank	Ankyrin repeat	PF00023	11
Adeno_knob	Adenoviral fibre protein (knob domain)	PF00541	11
Adeno_shaft	Adenoviral fibre protein (repeat/shaft region)	PF00608	11
rve	Integrase core domain	PF00665	11
BTB	BTB/POZ domain	PF00651	10
RNase_H	RNase H	PF00075	9
Sushi	Sushi repeat (SCR repeat)	PF00084	9

Table 1. For each Superkingdom are the ten most frequent PFAMs listed toghether with their PFAM Description and Accession number. 'Count' represents the number of appearances of the PFAM model in our data.