Published online 13 February 2019 *Nucleic Acids Research, 2019, Vol. 1, No. 1* ***1–1***
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 adjacently repeated amino acid sequence patterns. 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.

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. 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).

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.

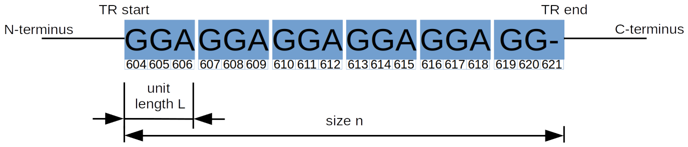


Figure 1. A sketch of a tandem repeat with it's descriptors. This micro TR with the ID A7TKR8 and a size of 6 units, each with a unit length of 3 amino acids shows a head-to-head orientation and consists of mixed units - a direct- and heterorepeat.

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).

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

[*] 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

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. TRs have been successfully exploited in bioengineering due to their design-ability (8). Despite much interest (9), the most recent and commonly cited census of protein TRs summarizing repeats from UniProtKB/Swiss-Prot protein knowledgebase dates back two decades (10). Since then the number of proteins in the curated protein databank SwissProt has grown more than seven fold (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. 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 IDP regions 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 the state of the art methods to annotate each protein with IDP regions and summarize the distribution of the overlap of TR and IDP regions over all kingdoms of life.

RESULTS

TRs are abundant in proteins of all domains of life

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. Therefore, we studied TR properties in four categories defined according to TR unit length L.

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 figure 5b and 5a respectively.

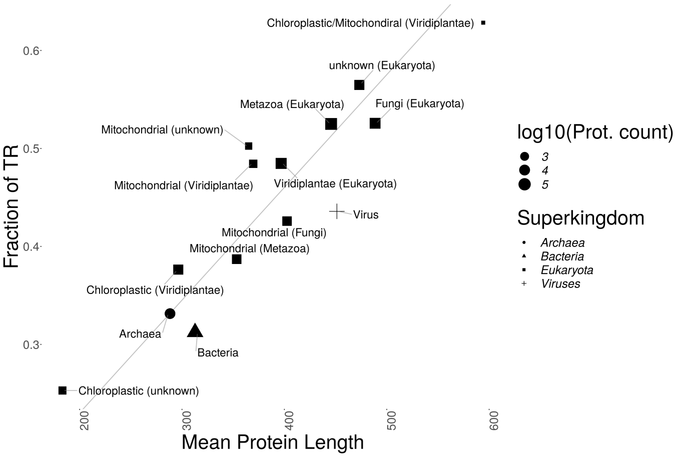


Figure 2. The fraction of proteins containing TRs over all protein entries in UniProtKB/Swiss-Prot is shown for each taxonomic Domain (Superkingdom) or Kingdom and displayed as function of the mean protein length and split according to the origin of the proteins. Chloroplastic proteins seem to be shorter and tend to have less TR than mitochondrial proteins. Non-mitochondrial and non-chloroplastic proteins appear to be longer and with more TRs.

Overall, 50.9% of all UniProtKB/Swiss-Prot 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%. A more detailed view is given in figure S2 where the percentage of proteins containing at least one TR is plotted against the average length of the proteins per species.

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 2 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. 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.

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.

or with?

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%) which is displayed in figure 3. Additionally, those small TRs also occur with many

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
micro TR fraction	0.117	0.109	0.245	0.191
short 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				
micro TR fraction	0.354	0.350	0.482	0.438
short 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, short TR prevail over the other types.

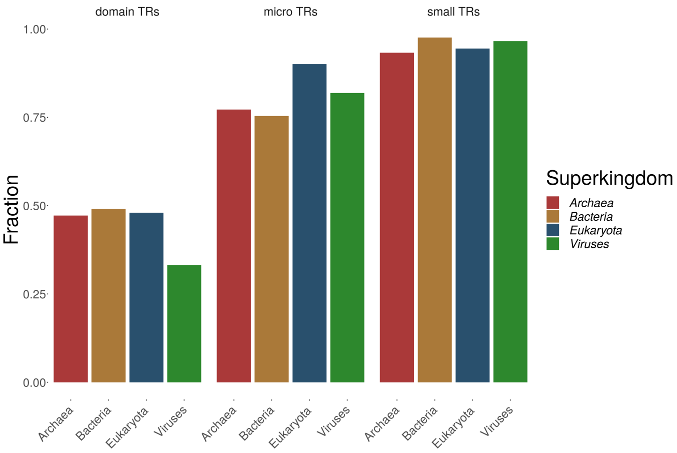


Figure 3. Proteins with ≥ 4 distinct TR regions are sorted by their TR type and shown kingdomwise. One can clearly see, that over all kingdoms small TRs dominate in proteins with many distinct regions.

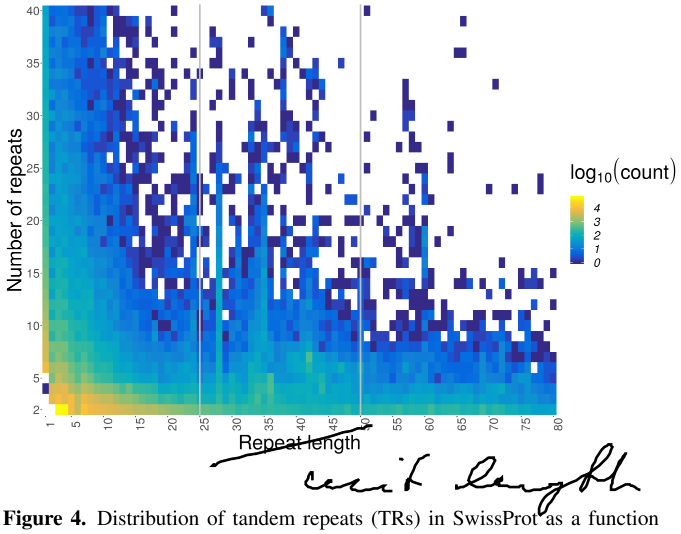


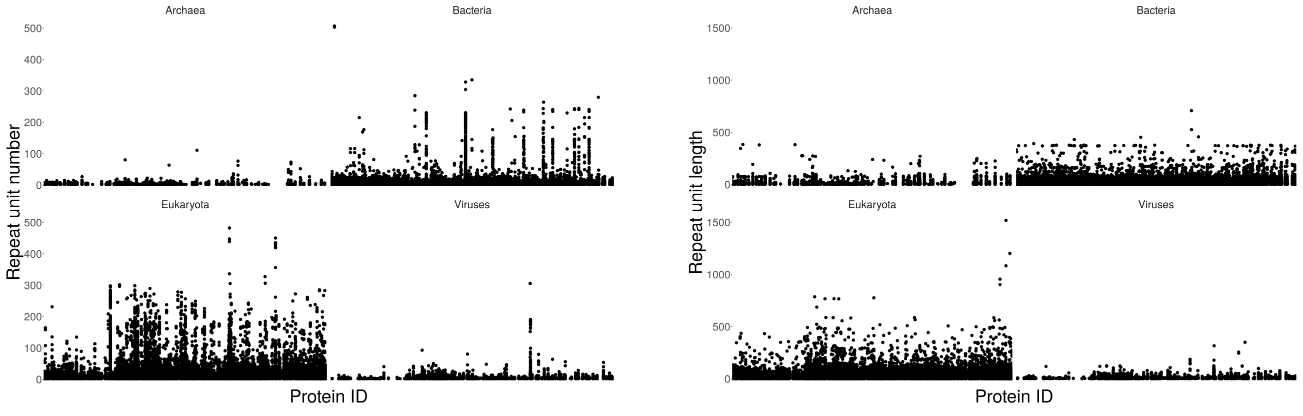
Figure 4. Distribution of tandem repeats (TRs) in SwissProt as a function of their repeat unit length $l_{effective}<=80$ (abscissa) and their number of repeat units $n_{effective}<=40$ (ordinate). Brighter colour indicates a larger number of TRs with a specific length and number of repeats. The majority of TRs has short 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 Zn finger TRs.)

repetitions. For example, multiple collagen and mucine related protein units fall in the segment of small TRs with high unit numbers. But also many *Staphylococcus* surface proteins and *S. epidermidis* protein such as (Q9KI14) containing 280 Serine-aspartate units or Notch1 proteins can be found in the same range.

Domain TRs mostly consist of few 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*.

Eukaryotic domain TRs tend to be more uniform distributed. 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 5a shows a peak of proteins containing many TR units. They seem to belong to collagen-like proteins of the Mimiviridae family.

In general, TRs are not homogenously distributed in terms of their unit lengths and numbers. Figure 4 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).



(a) Distribution of TRs by TR size.

(b) Distribution of TRs by TR unit length

Figure 5. 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.

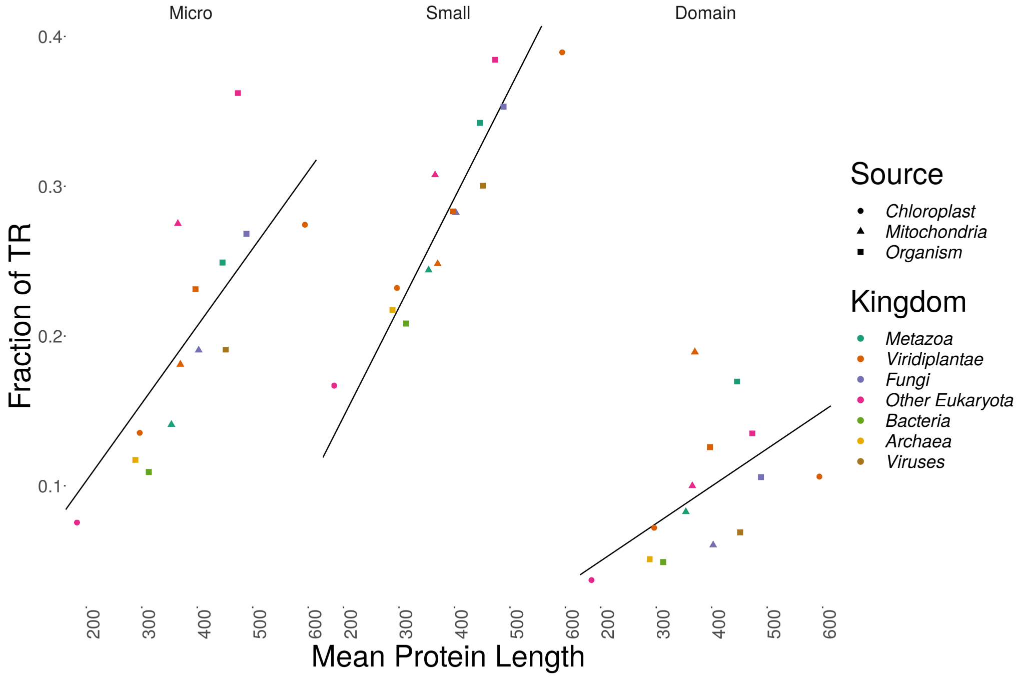
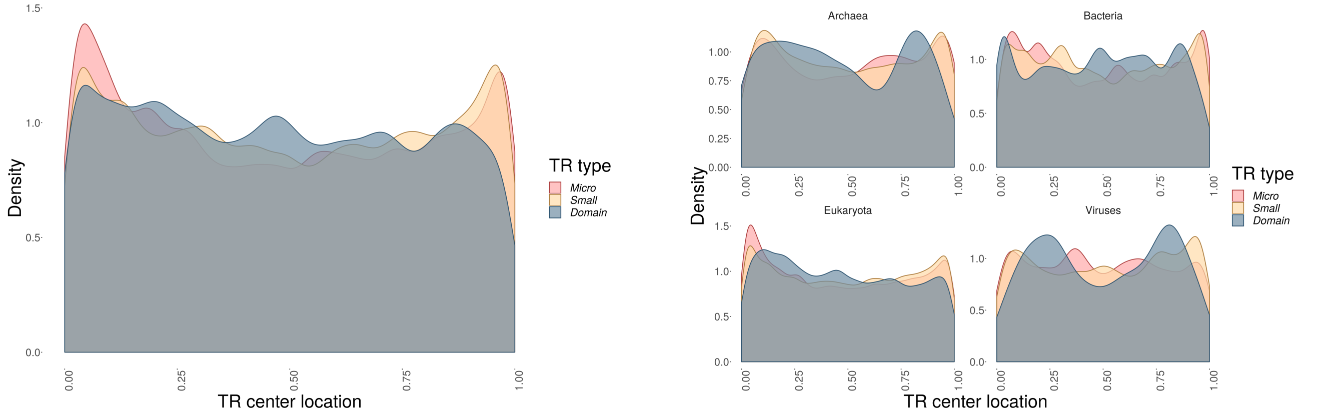


Figure 6. The amount of TRs (normalized by the amount of protein entries of the species) is displayed separately for each TR-type as a function of the mean length of the proteins. It can clearly be seen, that TRs appear mostly as small TRs. Comparing the fraction of TRs kingdom-wise, some clear tendencies 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.

It can be seen on plot 5b as one of the bacterial outliers 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.

More TRs are found in longer proteins

Figure 6 shows that in general, differences in TR distributions observed between kingdoms can be largely attributed to protein sequence length. Looking specifically at proteins which contain TRs, we can see that on average longer protein sequences tend to have more TRs. Indeed, we observe a strong linear relationship between the protein length and the fraction of proteins with TRs across all kingdoms of life for all TR types: $R^2 = 0.64$, p-value< 0.001 for micro TRs; $R^2 = 0.88$, p-value=< 0.001 for small TRs, and $R^2 = 0.24$, p-value=0.08



(a) Relative positions of TR within all proteins.

(b) Relative positions of TR within all proteins by Superkingdom.

Figure 7. 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.

for domain TRs (see also Supplementary S3, S4, S5). The relationship 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. Small TR seem to be the most recurrent in all kingdoms.

The TR-types are distributed in the same proportion for Prokaryotes (2 : 4 : 1 for micro:short:domain) and between Eukaryota and Viruses (1 : 3 :∼1 for micro:short:domain).

TR location is biased towards flanks for shorter TRs

Next, we explored where in a protein TRs 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 the different kingdoms and for different TR unit length categories.

As expected, TR relative position is shown to be prefered to the beginning of proteins. This seems to hold especially for domain TRs. However, shorter TRs (micro TRs and small TRs) displayed stronger preferences towards both, N- and C-terminals of SwissProt 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 Archaea show a similar behavior to be located towards both flanks of proteins, notwithstanding, our findings are based on a relatively limited number of observations (1290 and 955 respectively), the results from such analyses should thus be treated with caution.

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). TRs are involved in gene regulating mechanisms (34, 35). More complex organisms like Eukaryotes require more genes to be regulated. Therefore, it’s reasonable to find more TRs in eukaryotic organisms. 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.

More than half of the proteins whith 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. Looking at the number of repeating units, and the length of the unit, it could be shown, that they are generally non-uniform distributed.

We could further show, that there is a positive correlation between protein sequence length and the number of TR units. This is consistent with the previous observation by 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, we found no significance if the number of TR units is increased - as for domain TRs.

Small TRs are the most recurrent TR type across all taxonomic domains. Hence, it could conceivably be hypothesised that due to their small unit length, the multiplication of the TR unit leading to evolution of large, thermodynamically stable proteins might be facilitated (36). Whereas, domain TR are more expensive to multiply and difficult to lead to a thermodynamical stable configuration. Anyway their origins and roles in the protein function remain unclear and need to be explored.

TRs are located with a general tendency to the N-terminal end of protein sequences. However, micro- and small TRs cluster to both termini. This finding might be explained by

check literature for this throughly

the fact that domain TRs tend to be near the N-terminal end, hence they push the location-distribution of TRs in a protein sequence to the front.

NEXT STEPS

- Analyze homorepeats to a further extend. Do they follow the observed trends of domain, small, micro TRs?

- Make a Pfam clan analysis, analogous to figure 4 in Marcotte1999. This could reveal (more) homology between eukaryotic and prokaryotic repeat families, as more data is available today. And probably some new repeat families could be detected.

- Check each protein for intrisic disordered regions and if they overlap with TR regions. This would allow to test the hypothesis, that low complexity regions such as TRs are associated with rapid evolution and that some unstructered, low complexity regions form a special class of TRs without a specific structre.

- Finally split up Discussion and Conclusion sections to explicitly discuss all findings incl. their limitations.

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-Value=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 ($p-Value = 0.1$, $n_{effective} > 1.9$). 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-value 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-value=0.01).

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 (especially 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:

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

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).

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.

BIOPHYSICAL PROPERTIES IUPred and Globplot take a different approach and use biophysical properties of disordered protein sequences to predict disorder.

IUPred estimates the total pairwise interaction energy, based on a quadratic form in the amino acid composition of the protein, predicting the ability of residues to form rigid structures.

Globplot is focusing on shorter functional disorder inbetween structured domains and using propensities for amino acids to be in globular or non-globular states.

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**, 6781.
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**, 435445.
7. Nithianantharajah, Jess and Anthony J. Hannan (2007). Dynamic mutations as digital genetic modulators of brain development, function and dysfunction. *BioEssays*, **29**, 525535.
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. fig
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)**, 761772.
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.

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

SECTION 1

SECTION 2

annotate with marcotte and now

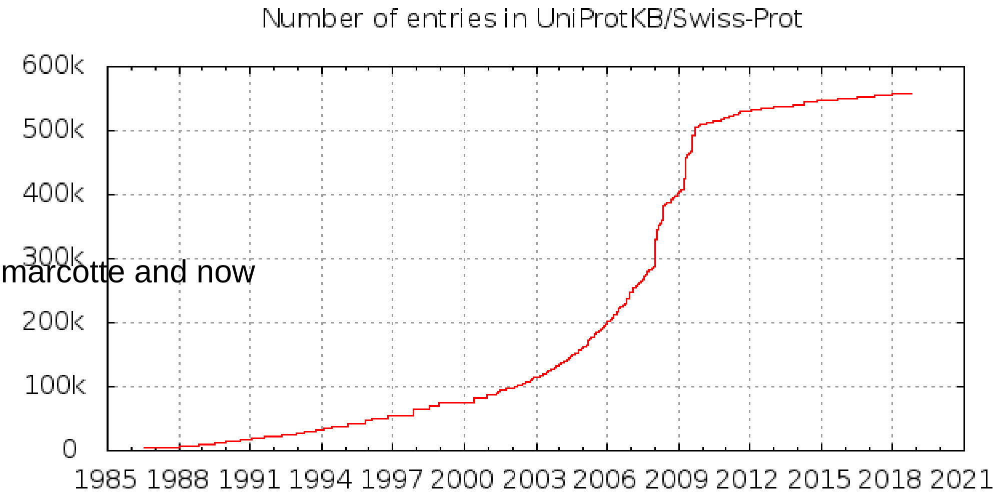


Figure S1. Summary of the growth of UniProtKB/Swiss-Prot protein knowledgebase. The last protein census dates back to the year 1999 (10). Since then, the entries in the UniProtKB/Swiss-Prot protein knowledgebase are grown more than seven fold. Figure from release 2018_09 statistics https://web.expasy.org/docs/relnotes/relstat.html,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

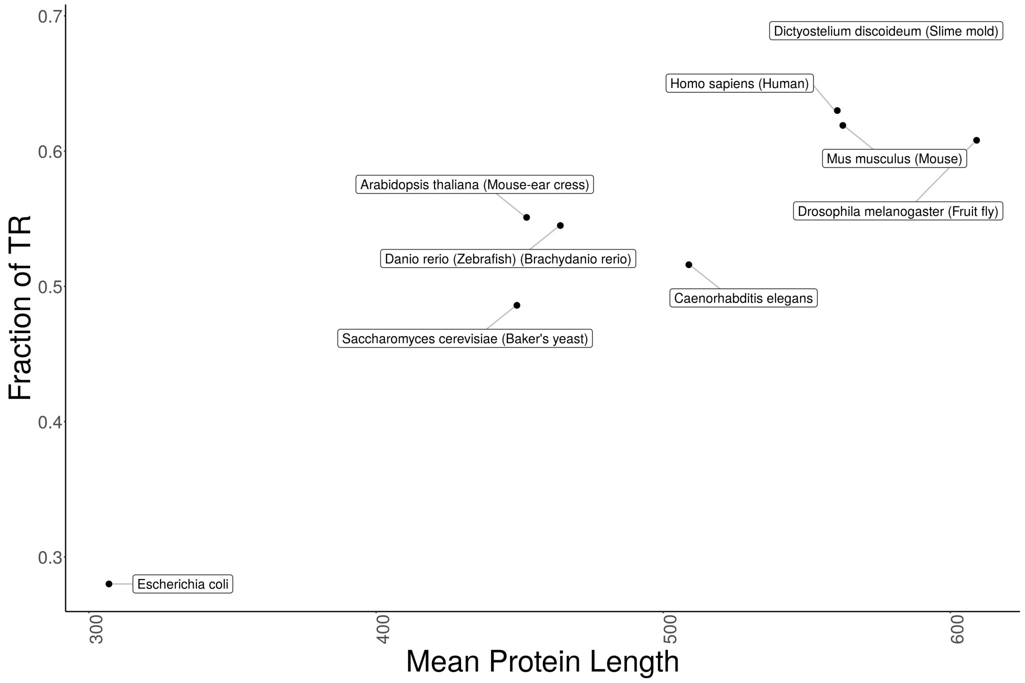


Figure S2. The fraction of proteins containing TRs over all protein entries in UniProtKB/Swiss-Prot is shown for a selection of species and displayed as function of the mean protein length.

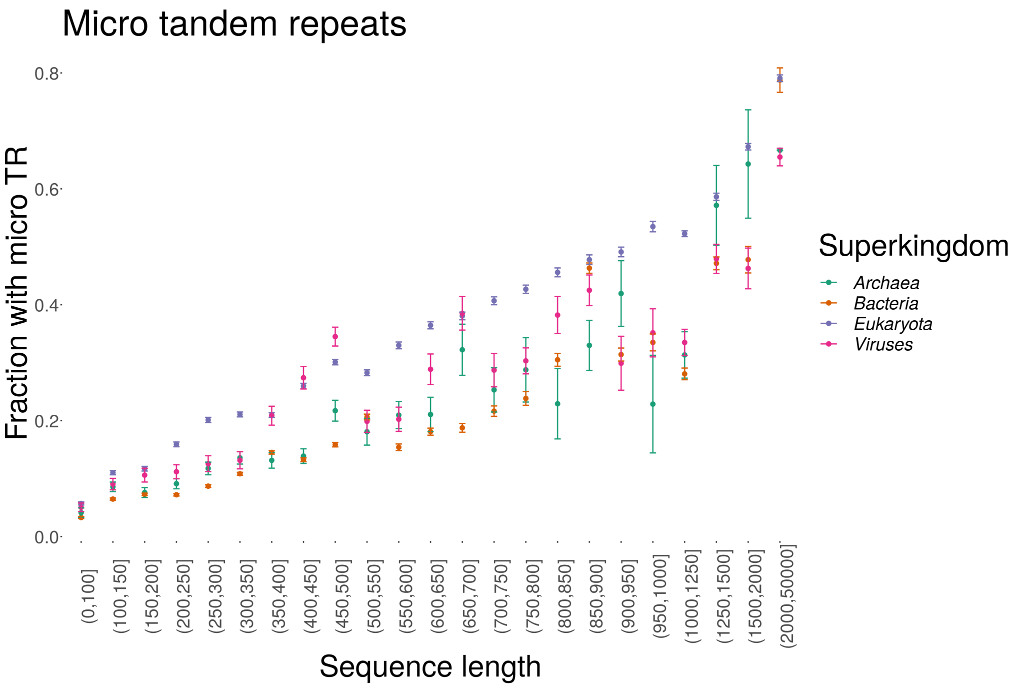


Figure S3. The fraction of proteins with micro TRs as a function of sequence length by kingdom resulting in a linear relationship.

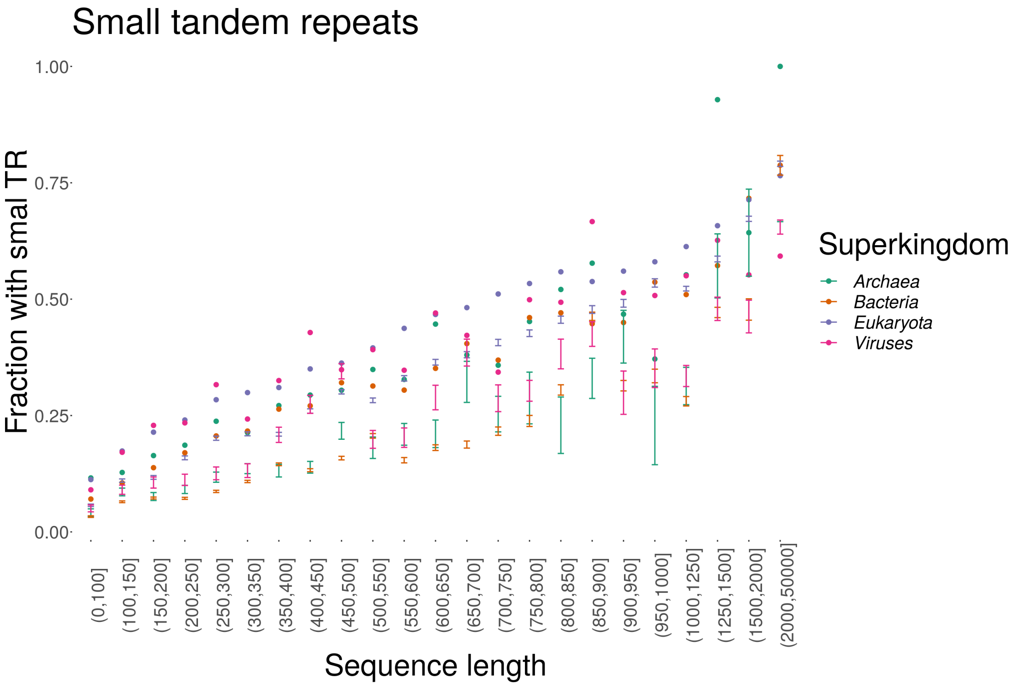


Figure S4. The fraction of proteins with small TRs as a function of sequence length by kingdom resulting in a linear relationship.

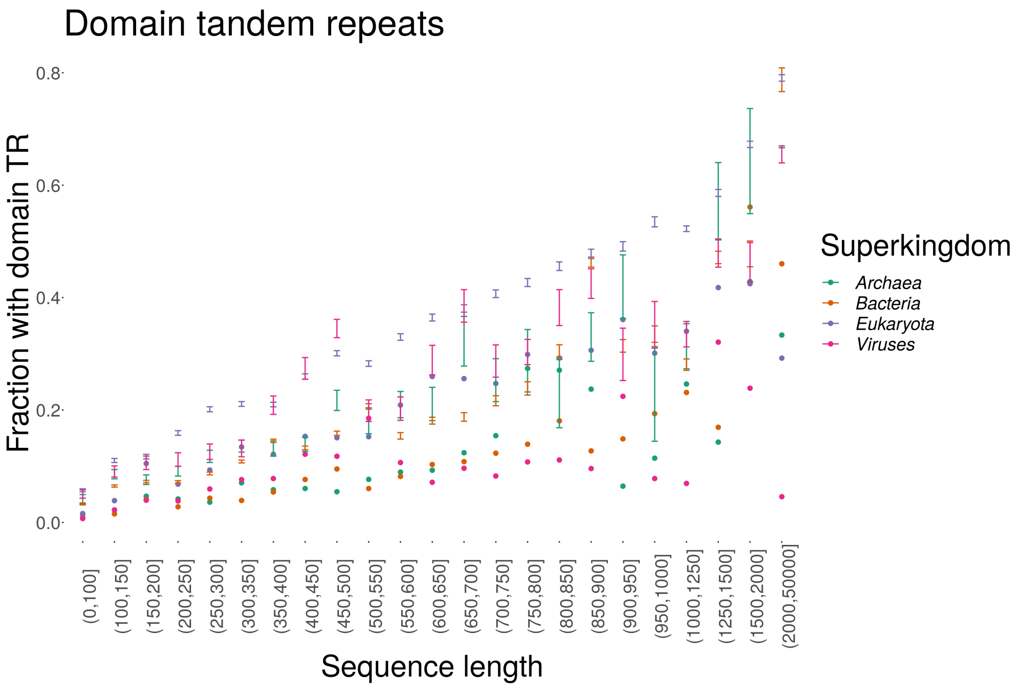


Figure S5. The fraction of proteins with domain TRs as a function of sequence length by kingdom resulting in a linear relationship.