Published online 31 January 2019

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

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

Maria Anisinova [1,*], Matteo Delucchi [2] and Second Co-Author [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

Some catchy abstract comes here.

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 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), 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 of 1 amino acid are herein called homo tandem repeats (homoTRs), protein TRs with 2-8 amino acids are herein called micro tandem repeats (micTRs) and mini tandem repeats (minTRs) for Protein TRs of a length of >9 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, dentatorubral-pallidoluysian atrophy, frontotemporal dementia, fragile X syndrome,

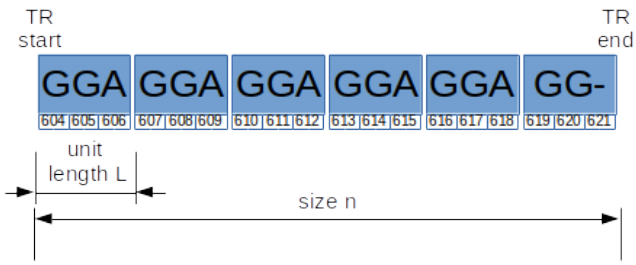

Figure 1. A sketch of a TR with it's descriptors. This micro tandem repeat with the ID A7TKR8 and a size of 6 units, each with a unit length of 3 shows a head-to-head orientation and consists of mixed units. Hence it is a direct- and heterorepeat.

fragile X tremor-ataxia syndrome, Huntington disease, spinobulbar muscular atrophy, spinocerebellar ataxia 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 which are part of 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 (REFs: Mar's paper and (5)), and that TR expansion is linked to the origin of novel genes (REF: Mar's papers). TRs have been successfully exploited in bioengineering due to

*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

their "design-ability" (8). Despite much interest (9), the most recent and commonly cited census of protein TRs summarizing repeats in UniProtKB/Swiss-Prot protein knowledgebase dates to nearly two decades ago (10). Since then the number of proteins in the curated protein databank Swiss-Prot 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 Swiss-Prot version 2015_11 (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 (REFs, (14)). 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 (e.g., (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.

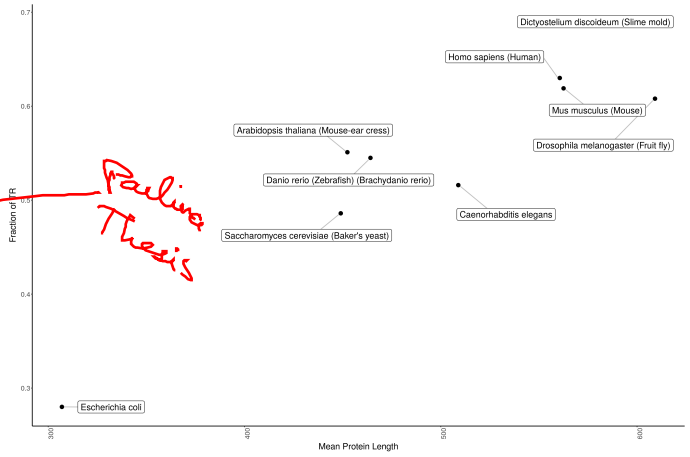

Figure 2. The fraction of proteins containing TRs over all protein entries in UniProtKB/Swiss-Prot is shown for a selection of species and displayed as function of the mean protein length.

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: (1) homorepeats L=1, (2) microrepeats 2<L<=3, (3) small repeats 4<L<=15, and (4) domain repeats L>15. Impressive numbers of TR annotations were predicted; their distributions with respect to TR unit length L and repeat number are summarized [kingdom-wise] in Table 1 and Figure 1. Overall, 50.9% of all UniProtKB/Swiss-Prot eukaryotic proteins contained at least one TR. In H. sapiens, 68.8% of all proteins contain TRs. Similar to M. musculus with 61.9% and D. melanogaster with 60.8%. In contrast stands E. coli with 28%. A more detailed few is given in figure 2 where the percentage is ploted against the average length of the proteins per species. 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 archaeic 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. By far the most frequent TRs were small repeats (XX% of all predicted TRs), followed by microrepeats (XX%), and domains (XX%). Homorepeats represent only XX% of all TRs. However, their proportion is still quite high, mostly due to eukaryotic proteins (8.5%). Short TRs also occur with high unit numbers. For example, a Staphylococcus epidermidis protein (Q9KI14) contains 280 Serine-aspartate units. Domain TRs mostly consist of few units. A prominent exception is an extracellular matrix-binding protein (Q5HFY8, Staphylococcus aureus) with 80 units each 97aa (PF07564) spanning 7700aa. In general, TRs are not homogenously distributed in terms of their unit lengths and numbers. 3 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 is zinc-finger (xx and xx aa), abundantly present as a TR in all domains of life, but also LRR (xx aa) and WD40-like beta propeller (39aa).

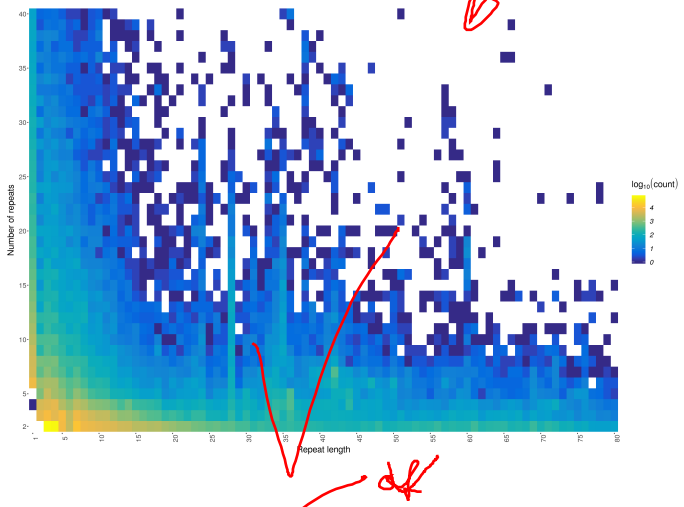

Figure 3. Distribution (Heatmap) of tandem repeats (TRs) in Swiss-Prot as a function their repeat unit length $l_{effective} \leq 80$ (x-Axis, x1) and their number of repeat units $n_{effective} \leq 40$ (x2, y-Axis). Darker colour indicates a larger number of TRs. The majority of TRs has short TR units. 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.)

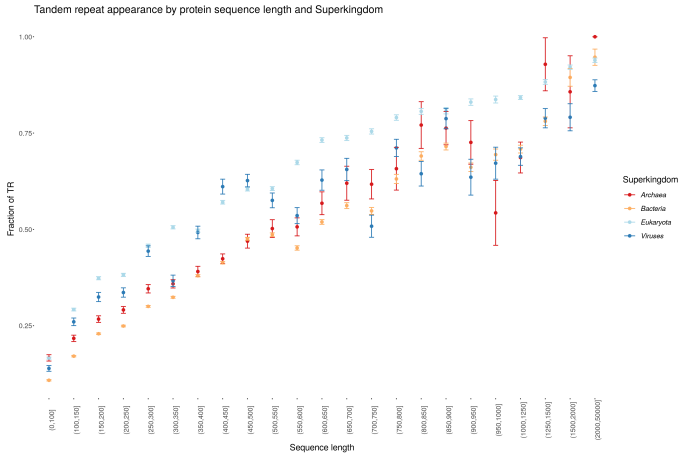

Figure 4. The fraction of proteins with TRs as a function of sequence length. [Note: which view is more helpful?]

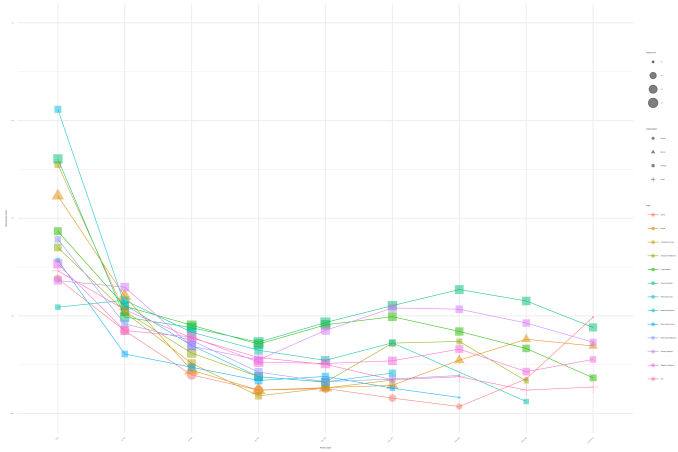

Figure 5. TMean disorder content as a function of sequnce length. [Note: how much complexity should we show?]

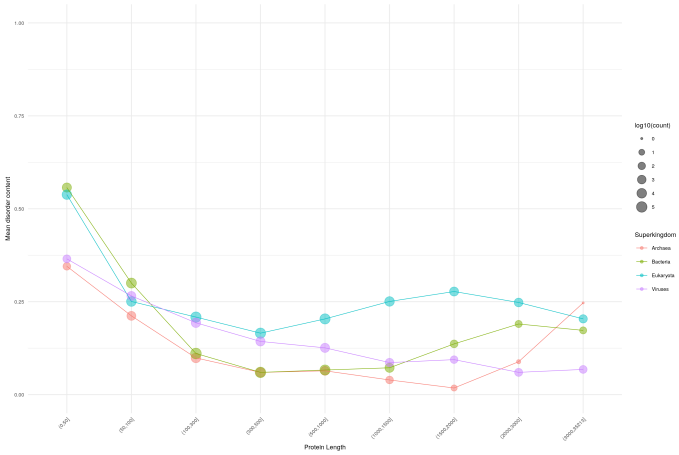

Figure 6. Mean disorder content as a function of sequnce length. [Note: how much complexity should we show?]

More TRs are found in longer proteins

In general, differences in TR distributions observed between kingdoms (Figure 1A-D ?) can be largely attributed to sequence length, with Eukaryotes having on average longer proteins. 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: $R^2 = XX$, p-value=XX for micro TRs; R^2 = XX, p-value=XX for small TRs, and R^2 = XX, p-value=XX for domain TRs. 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. 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 5a. PolyQ and polyN homorepeats may often be observed with > 50 repetitions. The 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. [Can we back this up or reject? for example, can we compare the extent of differences of protein lengths with polyN and polyQ (separately) for eukaryotes vs. other?] The observed positive correlation between protein length and TR quantity is consistent with the previous observation by Marcotte et al (1999). However, our study also warns against over-generalization of such observations. Clearly, protein sequences are highly heterogeneous in their origin, content, structure and

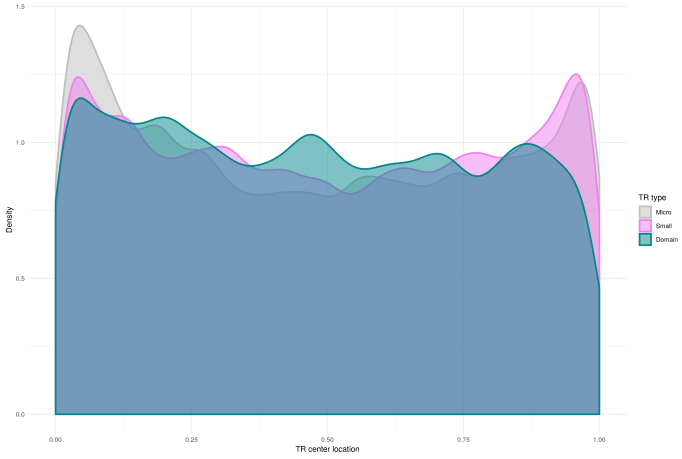

Figure 7. Density plots for the relative positions of tandem repeats (TRs) within the protein for four Superkindoms. Colours indicate repeat unit lengths. Interestingly, short TRs are biased towards the flanks of the protein. In particular for Eukaryotes, there is a clear correlation between TR unit 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, Bacteria show a strong bias to the C-terminal protein flank. [Note: I renormalization seems important.]

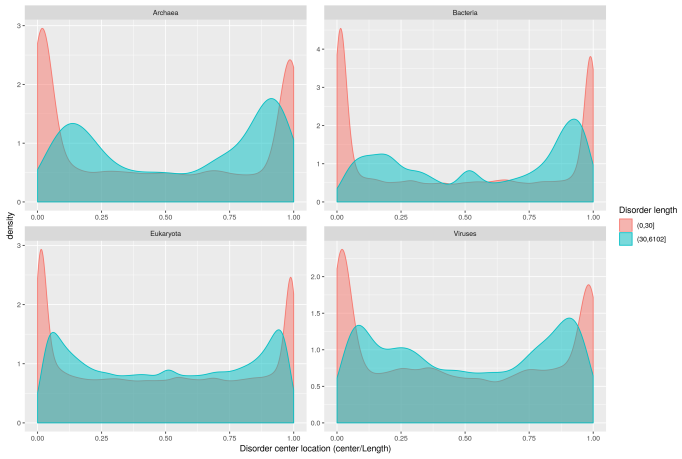

Figure 8. Density plots of position of disorder regions within the protein for four Superkingdoms. Both short and long disorder regions tend to cluster towards the flank of the protein, to the N-terminal specifically, with the trend being somewhat weaker in Eukaryotes.

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, [... Here, exceptions from the rule should also be mentioned if any striking ones are observed. This satisfies Arne's concerns.] [Figure 2b: can we add anything to above based on this figure?]

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). [does it make sense to also do these plots for the starting position. would this be interesting?] The observed distribution of TRs along the protein length was non-uniform and dependent on the TR unit length. Figure 3a shows the distributions of the relative positions of TRs in proteins across the different kingdoms and for different TR unit length categories [can we have only 4 categories here, as above?]. As expected, domain TRs were typically centered around the middle of a protein. However, shorter TRs displayed stronger preferences towards N- and C- terminals of Swiss-Prot proteins. In particular for Eukaryotes, there was a clear correlation between the TR unit length and the location bias towards the protein flanks. In eukaryotic proteins TRs were overrepresented in the N-terminal protein flank, while in Archaea and Bacteria, the TR preference was towards the C-terminal. For homorepeats such tendency was particularly striking, particularly in Archaea, where most homorepeats were found in the C-terminal. [Again any special cases could be mentioned here. For example in viruses we see a blue spike towards the C-terminal. Does this correspond to any special features of viruses also see in the figure 1, for some specific TR?] [ToDo: Potentially compare to random distribution. Or, normalize by "first possible occurrence", instead of "sequence length".]

rewrite, as this was written before normalization

TR clustering analysis

Around 58% of short and micro repeats, and 63% of domain repeats are found in proteins annotated with a Pfam clan.

TODO: [Plots and descriptions of most prevalent Pfam clans]

Most TRs appear to be disordered

TODO: Add Table on TR-IDP overlap TODO: Add MCC info

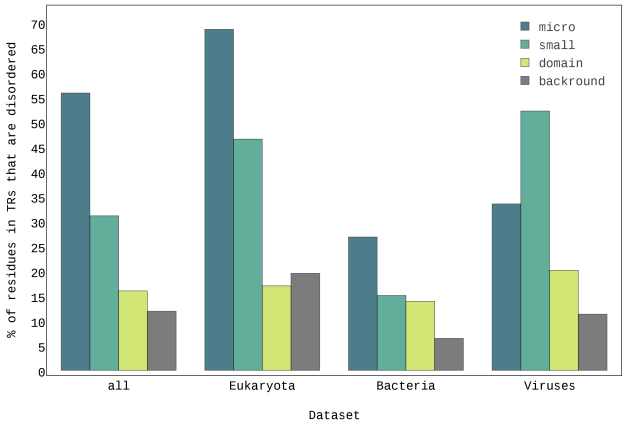

Figure 9. % of disordered residues in TRs, split by unit size and kingdom

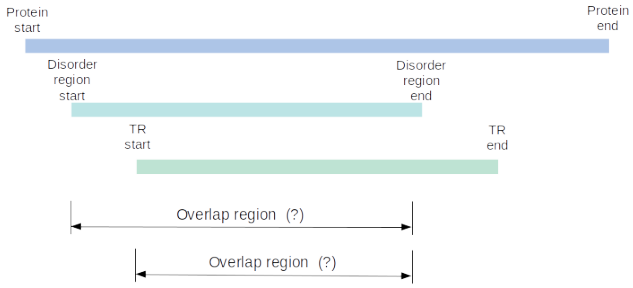

Figure 10. A sketch of a overlap region in the context of a protein with TRs.

TRs with small unit sizes are most disordered

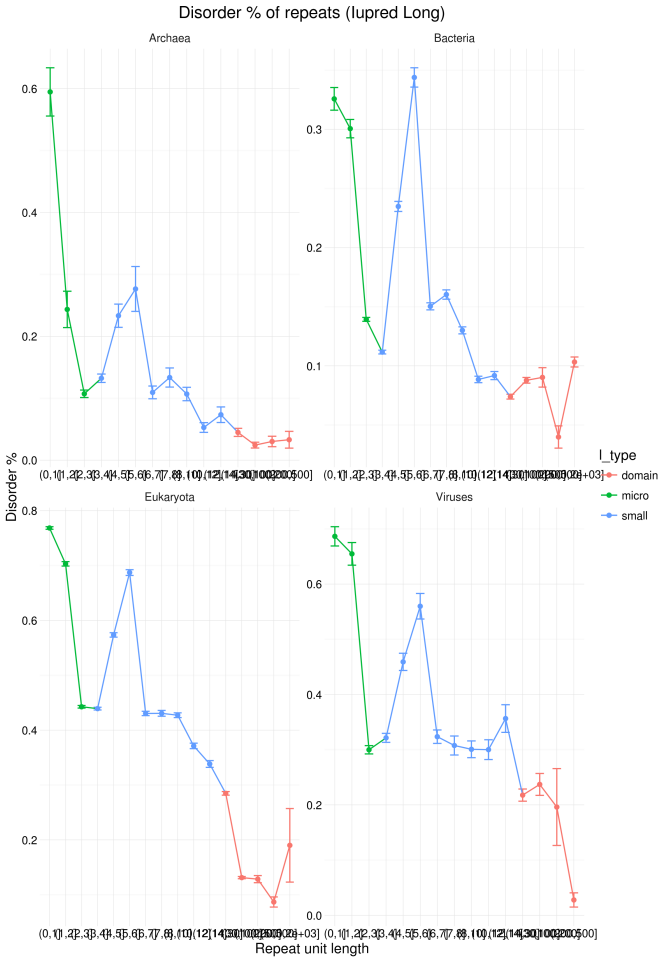

Figure 11. % of disorder within a TR for different unit lengths

TR and disorder overlap is mostly explained by skewed amino acid frequencies

Predictions are not always correct Figure X: Logo of selected unit sizes, like collagens, explaining how the same amino acid distributions can end up being predicted disordered when they are in fact part of coiled coil, beads on a string and other known structures.

TODO: Create final version of logos of unit size 6

DISCUSSION

Summarize obtained results, discuss in the context of existing literature.

[TR numbers are consistent with our recent estimates for human/plants. Numbers are even higher than previously estimated (Marcotte paper), due to much better annotation, taking all TRs (short, long) into account. Short TRs dominate TR landscape but are mostly uncharacterized. Their origins and roles in the protein function remain unclear and need to be explored.]

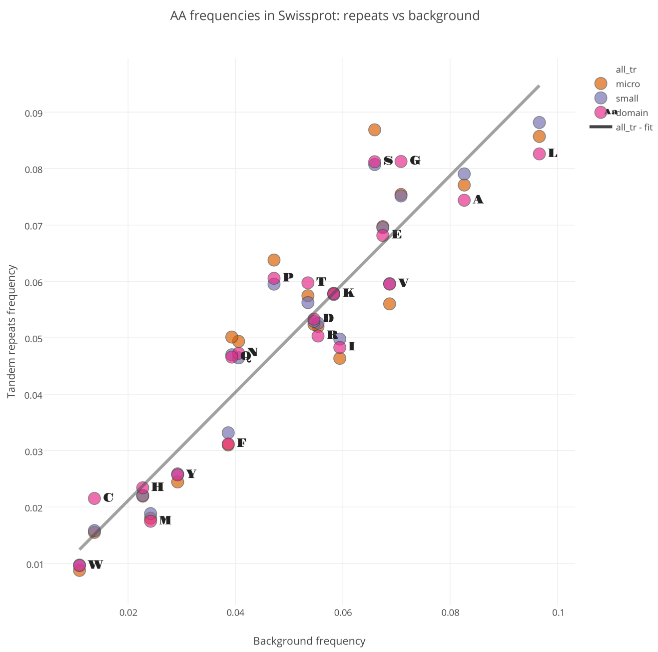

Figure 12. AA frequencies in repeats vs background

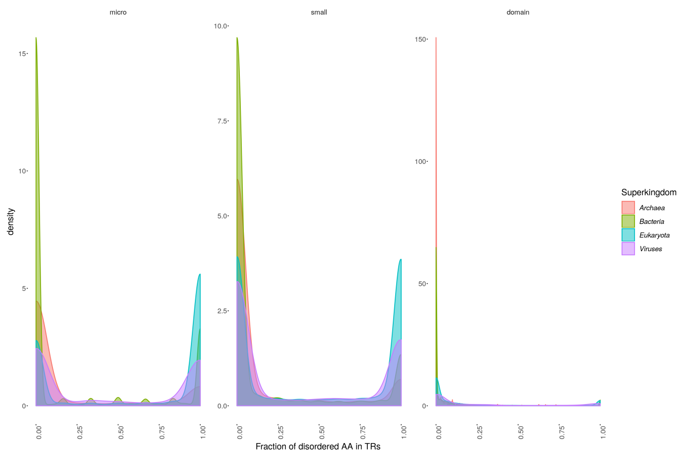

Figure 13. Shown is the fraction of disordered chars for micro TRs, small TRs and domain TRs. We see clearly bimodal distributions, with the majority TRs mostly ordered, however with a striking fraction of small TRs and micro TRs mostly disordered.

Localization of TRs and disorder.

Functional relevance of interesting categories

Discuss most frequent PFAM domains for different kingdoms and their functions.

Hypothesize about evolutionary and functional mechanisms of repeats going together with disorder.

E.g., discuss hypothesis that most TRs evolved after the split of eu and prokaryotes (Marcotte.) E.g., we could

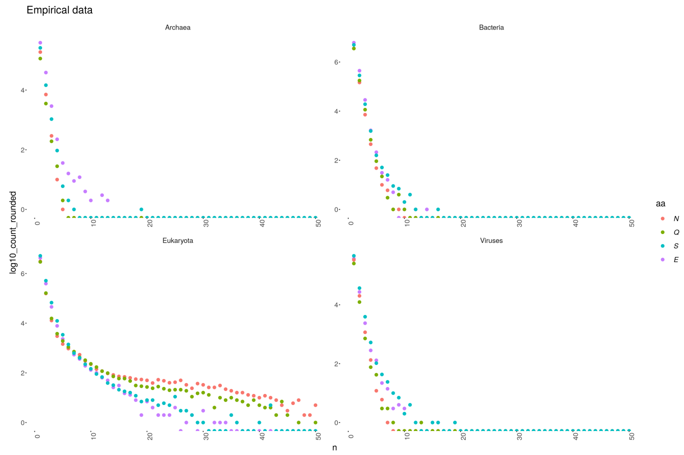

Figure 14. Count of homorepeats in Swiss-Prot in four Superkingdoms for different repeat unit number $(n<=50$, equivalent to repeat length) for amino acids E, S, N and Q. Homorepeats with large n seem to mostly pertain to the Eukaryotes.

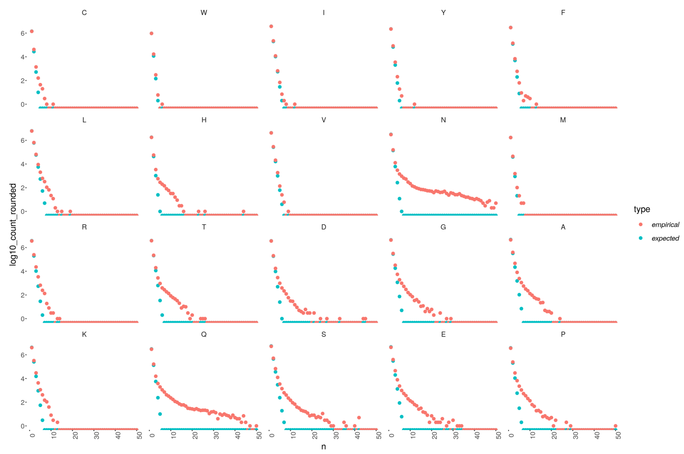

Figure 15. Empirical and expected count of homorepeats in Swiss-Prot Eukaryotes $(n<=50)$. Amino acids are ordered by their propensity to promote structural order.

connect disorder annotations with ”conserved/diverged” TR annotations in human/plants. E.g., comment what we could learn from an evolutionary study.

CONCLUSION

Text. Text.

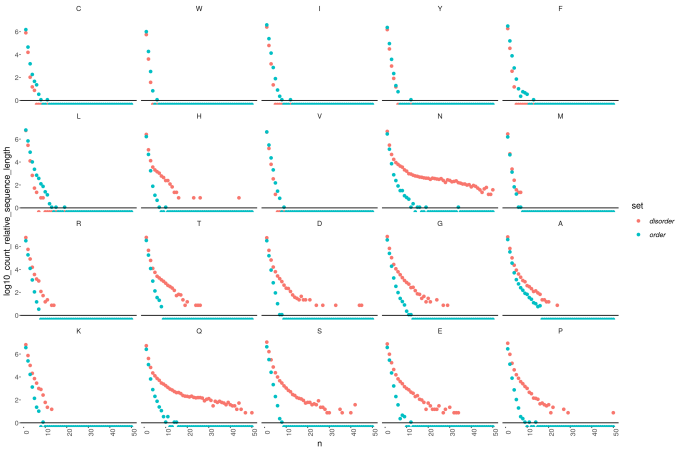

Figure 16. Empirical count of homorepeats in Swiss-Prot Eukaryotes ($n <= 50$) for ordered and disordered regions (consensus MobiDB annotations, no minimum length cut-off.). Amino acids are ordered by their propensity to promote structural order.

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

MATERIAL & METHODS

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 crystallise, 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 – *Jul* – 2014)) . MobiDB provides consensus annotations as well as raw data from DisProt, PDB (missing residues in X-Ray and NMR) and 10 computational predictors.

[Currently running Disopred]

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), definiing 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.

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 $l <= 3$ as micro TRs, as they correspond to nucleic microsatellites. Further, we denote TRs with $3 < l < 15$ as small TRs, and TR with $l >= 15$ as domain TRs.

Statistical significance filter The shorter and the more diverged a TR, the harder it is to distinguish from none-TR sequence. 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 Jorda:19671691, XSTREAM

Newman:17931424 and HHrepID Biegert:18245125 (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 Swiss-Prot. The corresponding sequence profile models were retrieved from PFAM Finn:26673716, 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).

Homorepeat annotations To compare expected and empirical number of homorepeats in Swiss-Prot, we exactly counted the number of runs of all lengths and all amino acids in Swiss-Prot. We repeated this exact count for bounded subsets of Swiss-Prots, 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., [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.] (http://www.sciencedirect.com/science/article/pii/S0898122110009284).

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

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 Opi

ACKNOWLEDGEMENTS

nion 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.

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

SECTION 1

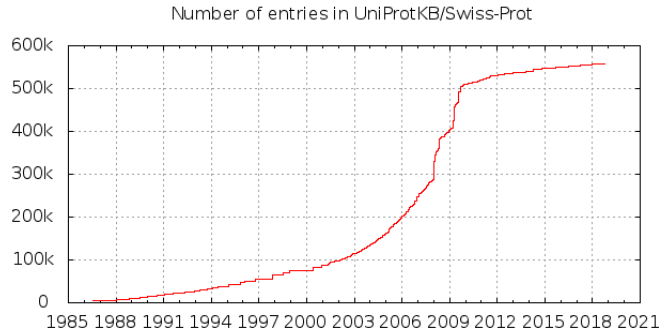

Figure S1. Summary of the growth of UniProtKB/Swiss-Prot protein knowledgebase. The last protein census dates back to the year 1999 (10). Since then, the entries in the UniProtKB/Swiss-Prot protein knowledgebase are grown more than seven fold. Figure from release 2018_09 statistics https://web.expasy.org/docs/relnotes/relstat.html,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