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

For citations use format: (First author:PUBMEDID)

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 sequence patterns, known as tandem repeats (TRs): These are highly abundant (e.g., in human proteome >60% proteins contain TRs), display an impressive variability of sizes, structures and functions ({Schaper:22923522},{Schaper:24497029}), have enhanced binding properties ( ) and are known to have associations with disease and immunity related functions ( ). 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 {Schaper:24497029}), 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 their "design-ability" (e.g., {Javadi:23831287}).

Despite much interest (eg, {Di Domenico:24311564}), the most recent and commonly cited census of protein TRs summarizing repeats in Swiss-Prot dates to nearly two decades ago ({Marcotte:10512723}). Since then the number of proteins in the curated protein databank Swiss-Prot has grown more than 7 fold (CHECK). Equally, a multitude of new methods were developed for the prediction and analysis of TRs (ee {Anisimova:25853125}). 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 ({Anisimova:25853125}{Schaper:25987568}).

Here we apply this recent methodology to characterize the distribution of protein TRs as currently found in Swiss-Prot version XXX (Boutet:26519399). 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.

release 2014_01 or 2015_11 (workflow_data_aggregation.md)

Further teins with TRs tend to be enriched with intrinsic protein disorder (IDP) ( ), and vice versa (REFs, {Szalkowski: 23877246}). Both TR and IPD regions also tend to be overrepresented in the hubs of protein–protein interaction networks (Ekman et al. 2006). 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

intrinsic protein disorder: is a protein that lacks a fixed or ordered three-dimensional structure. IDPs cover a spectrum of states from fully unstructured to partially structured

have "beads on a string" organizations ({Kajava:21884799}). But there is undoubtedly a class of protein TRs strongly associated with unstructured regions (e.g., {Jorda:20553501} {Szalkowski: 23877246}). Several studies have shown that compositionally biased, low complexity regions, often found in IDPs evolve rapidly, including recombinatorial repeat expansion events ({Tompa:12938174}, {Simon:19486509}). Others in contrast observed that the association between repeat enrichment and protein disorder is not as clear [Light et al. 2013]. 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

Schaper, Elke et al. (2015).
"TRAL: Tandem repeat annotation library". In: Bioinformatics
31.18, pp. 3051–3053. issn: 14602059. doi: 10.1093/bioinformatics/btv306

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 (REF); 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 215. [Please check throughout!] 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. [Need 4 colour matrix figures – for each kingdom and viruses. We will decide what to show in main text and what can be presented in supplement.] Overall, 59.3% of all UniProtKB/Swiss-Prot eukaryotic proteins contained at least one TR. [Were majority in any specific organism? Or simply give examples. For example, this was XX in human, XX in drosophila and XX in yeast. There were interesting examples in Jorda & Kajava 2010, although just for homorepeats.] Interestingly, 52.4% of viral proteins contained TRs, almost as frequently as in eukaryotes. In comparison, fewer prokaryotic proteins contained TR, but nevertheless >40% for both bacterial and archaeic proteins. A substantial fraction of proteins contained more than one distinct TR region, most frequently in eukaryotic proteins (XX% of all proteins with TRs), but also in viral (XX) and prokaryotic proteins (XX in Bacteria and XX in Achaea). In eukaryotes, XX% (XX absolute count) of all proteins with TRs had 4 (or more) distinct TR regions. [Any of these are worth mentioning?] 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%). [Human homorepeats are well known for their disease associations.] Short TRs also occur with high unit numbers. For

(1) small

necessary to mention?

def of small?

example, a Staphylococcus epidermidis protein (Q9KI14) contains ~280 Serine-aspartate units. [Check] 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. Figure 1 reveals multiple peaks, showing that some unit lengths are particularly frequent. These peaks represent common TRs, with specific TR units used in varying number. [Mention some striking outliers by kingdom: first common feature for all kingdom and then outliers specific to each kingdom and viruses separately.] 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). [In bacteria, a common TR is the bacterial transferase hexapeptide (36aa), closely followed by ?? In viruses, ???]. [Interesting differences between kingdoms need to be described here, if any.]

definition? just by eye or stats? REF?

More TRs are found in longer proteins

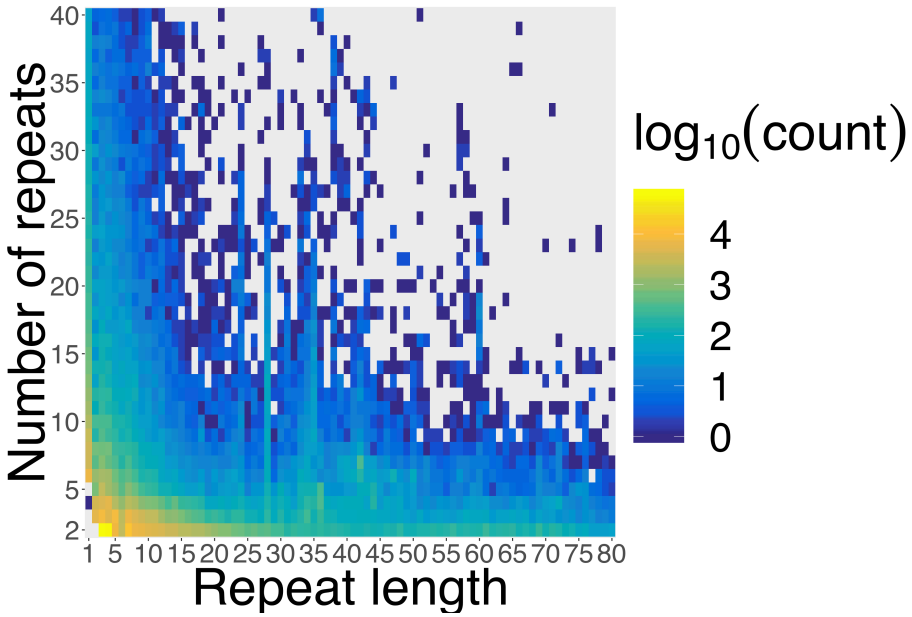


Figure 1:

Fig. 1a: Distribution (Heatmap) of tandem repeats (TRs) in Swiss-Prot as a function their repeat unit length $l_{effective}$ <= 80 (x-Axis, x1) and their number of repeat units $n_{effective}$ <= 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.)

bad format

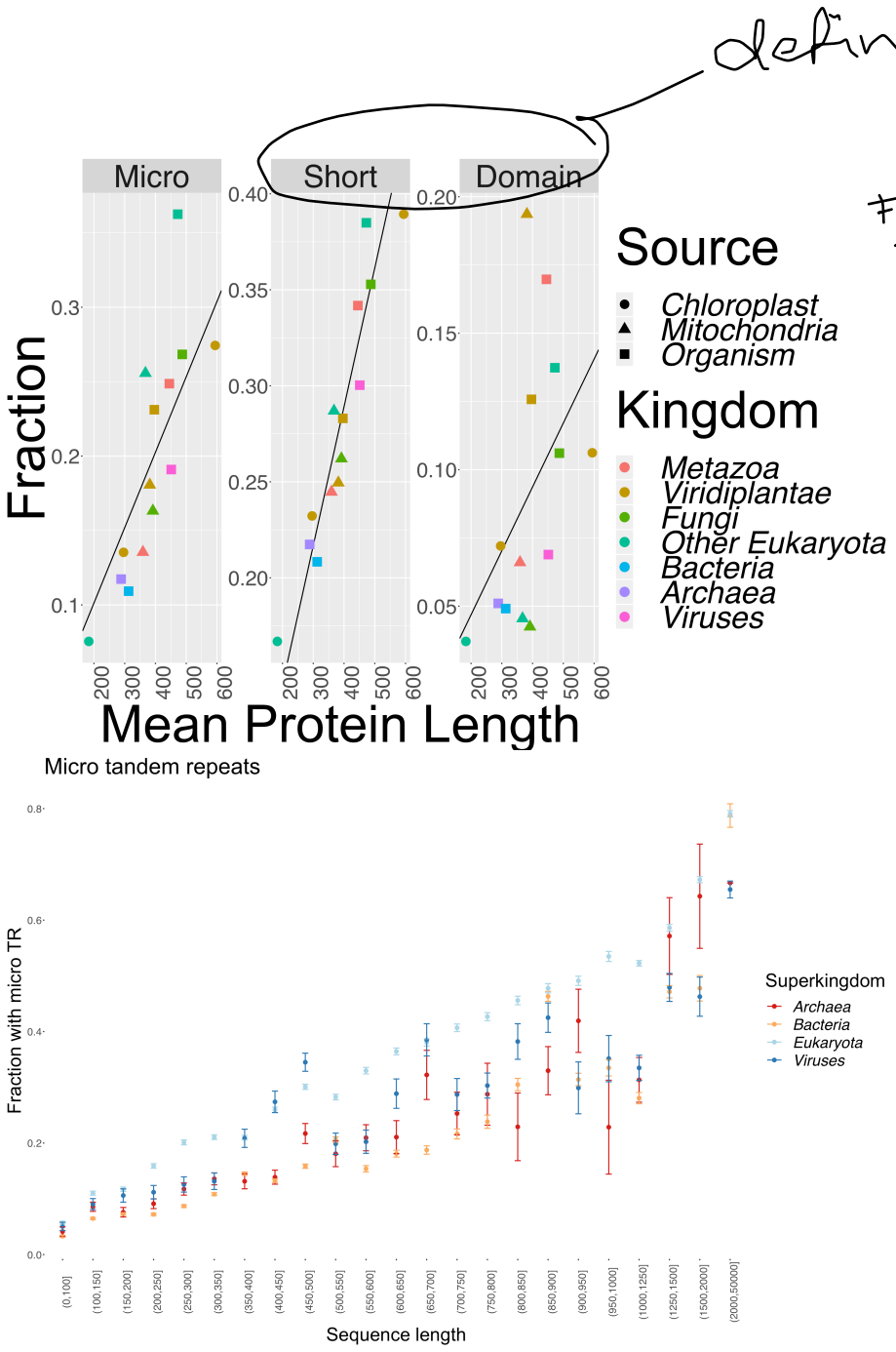


Fig. 2a) The fraction of proteins with TRs as a function of sequence length. [Note: which view is more helpful?]

2nd

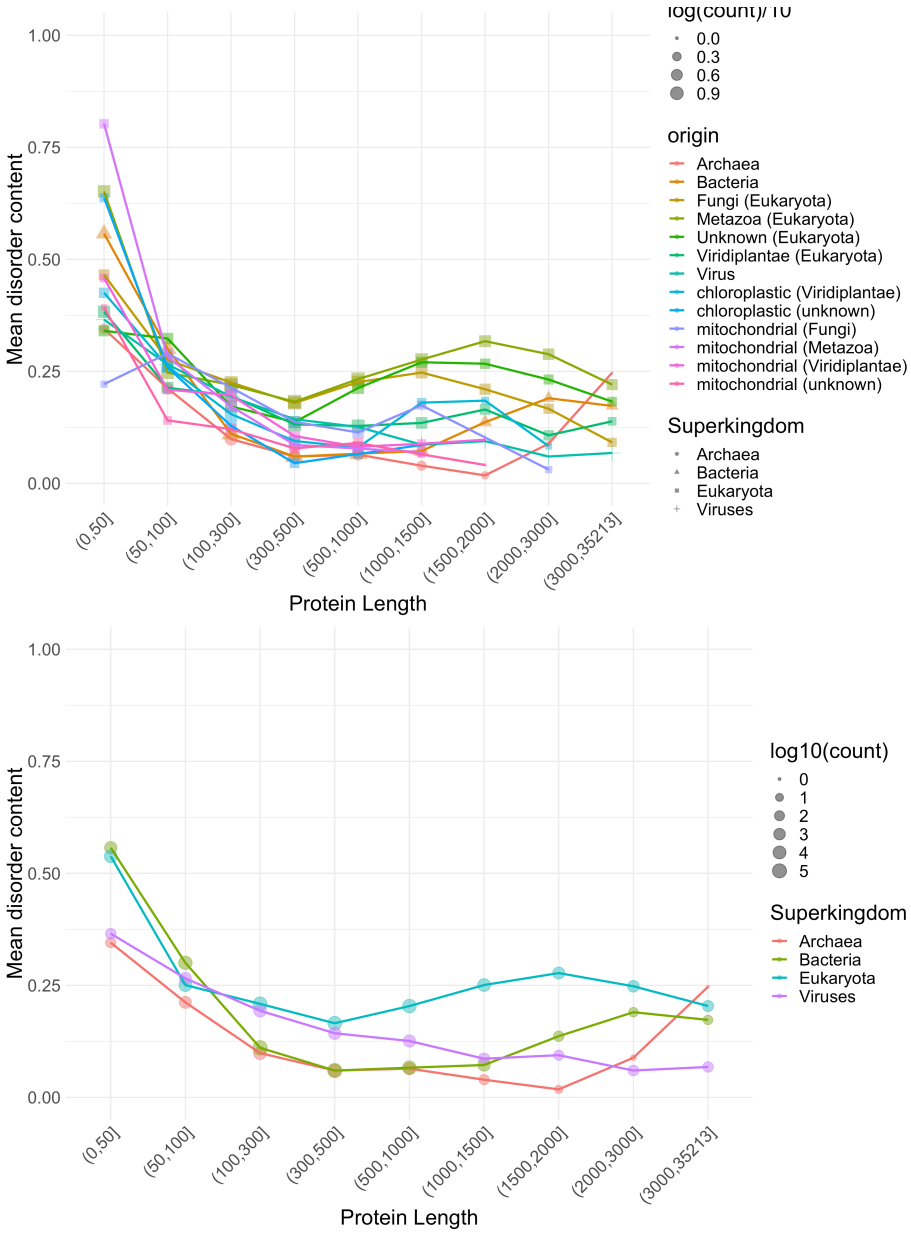
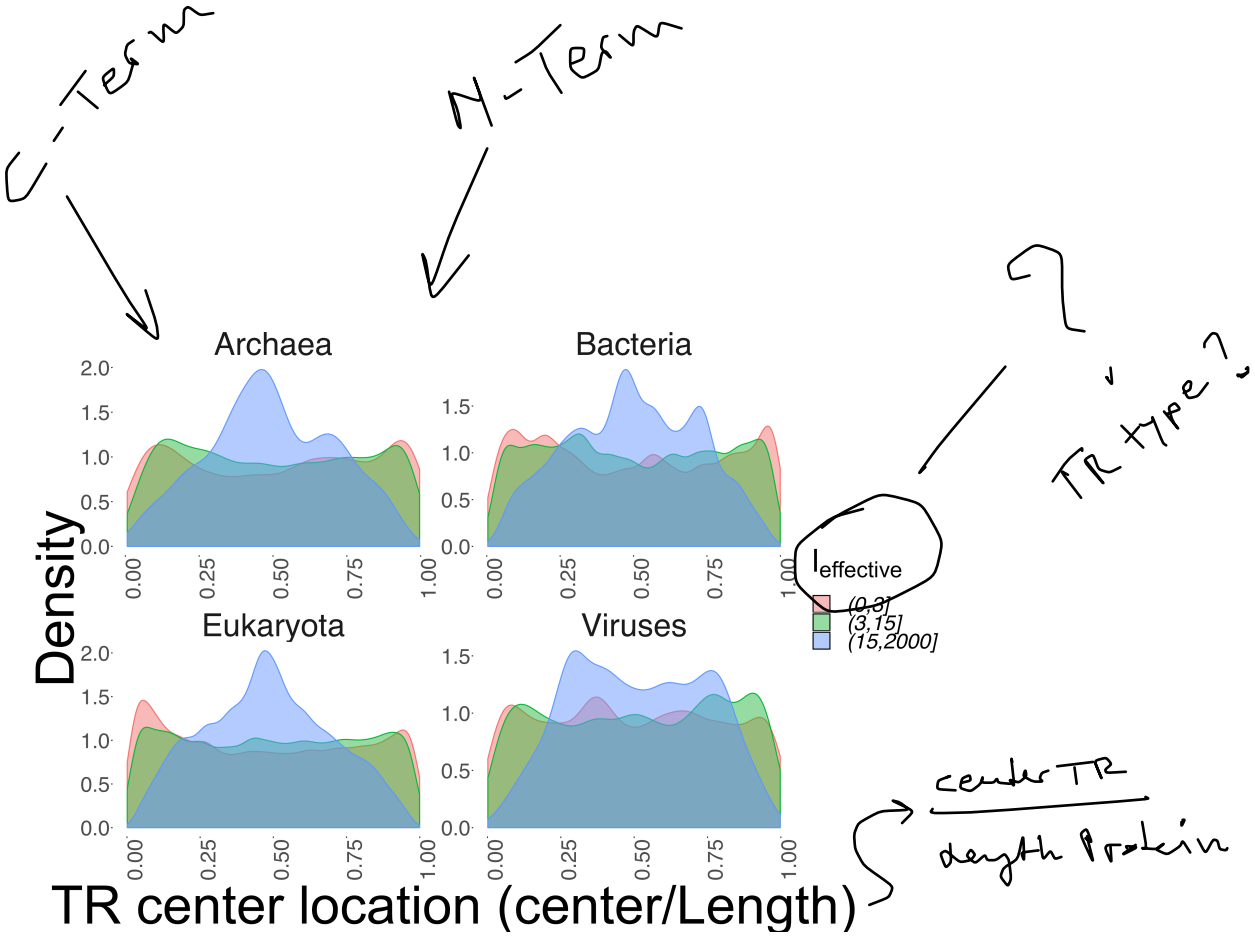


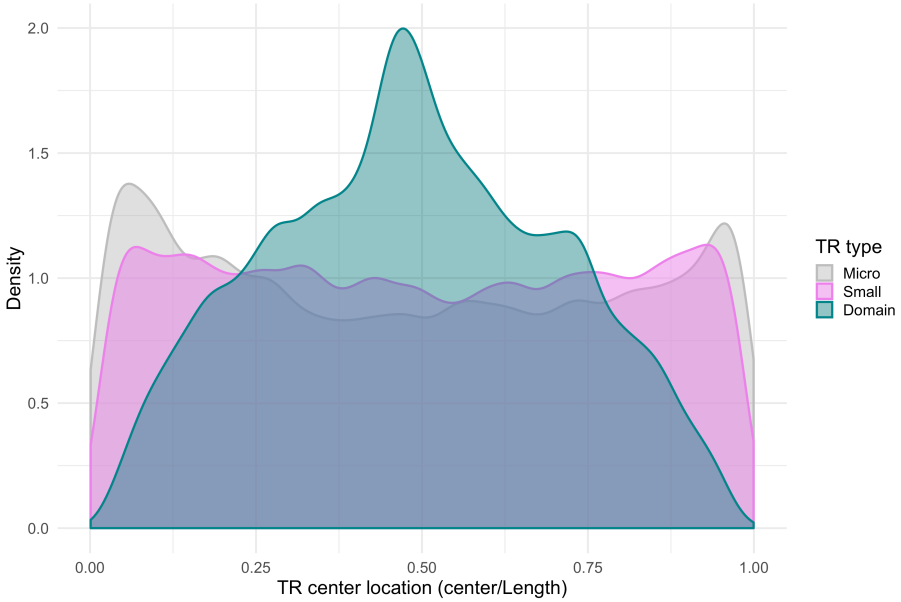
Fig 2b. Mean disorder content as a function of sequnce length. [Note: how much complexity should we show?]

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 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


C-Term
N-Term
TR type?
Archaea
Bacteria
Eukaryota
Viruses
Density
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
l effective
(0,3]
(3,15]
(15,2000]
center TR
length Protein
TR center location (center/Length)


Density
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
TR center location (center/Length)
TR type
Micro
Small
Domain

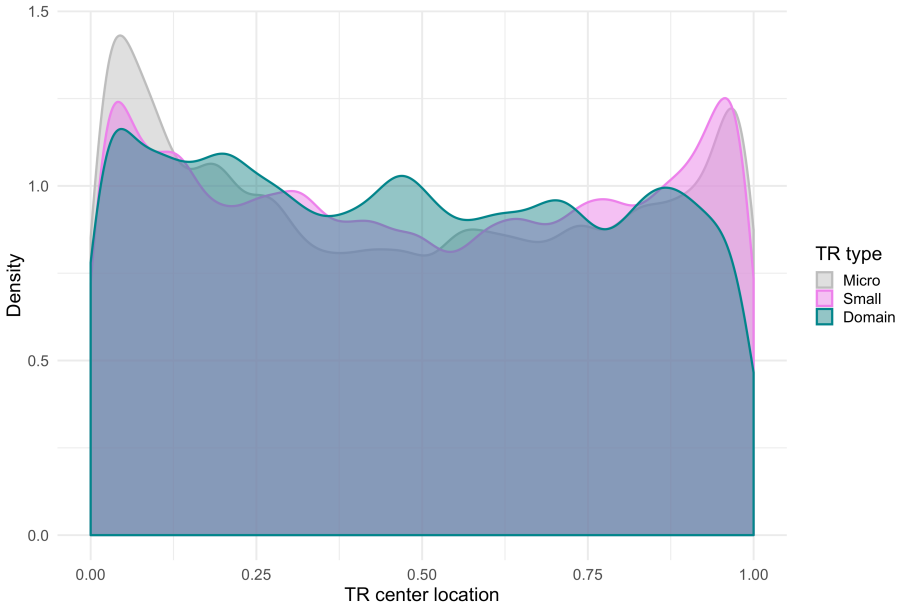


Fig. 3a. Density plots for the relative positions of tandem repeats (TRs) within the protein for four Superkindoms. Colours indicate repeat unit lengths. Interestingly, short TRs are biased towards the flanks of the protein. In particular for Eukaryotes, there is a clear correlation between TR unit length 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.]

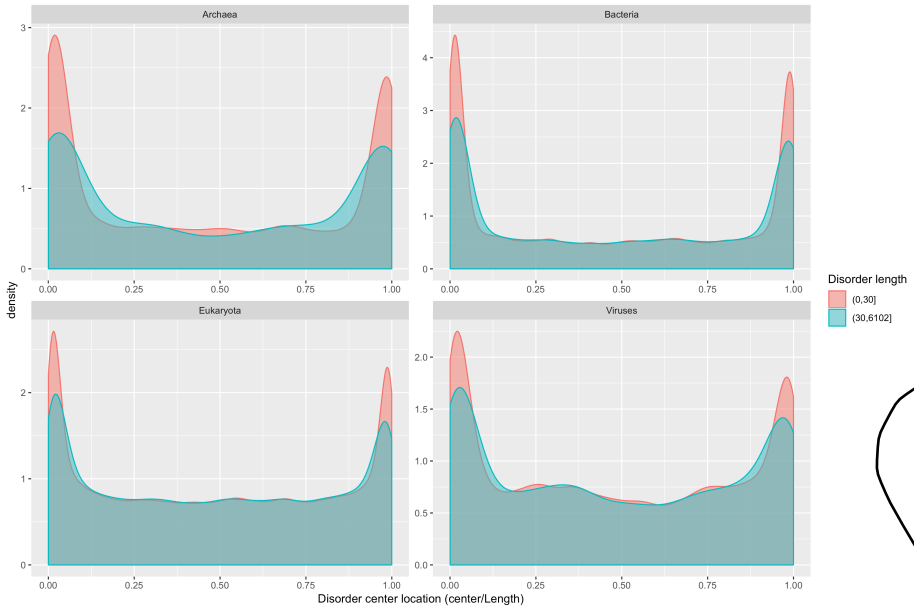


Fig 3b. Density plots of position of disorder regions within the protein for four Superkingdoms. Both short and long disorder regions tend to cluster towards the flank of the protein, to the N-terminal specifically, with the trend being somewhat weaker in Eukaryotes.

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".]

normalize by TR length. Evcar. have longest TR & the most (?) Hmm it's logic to have them uniform

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

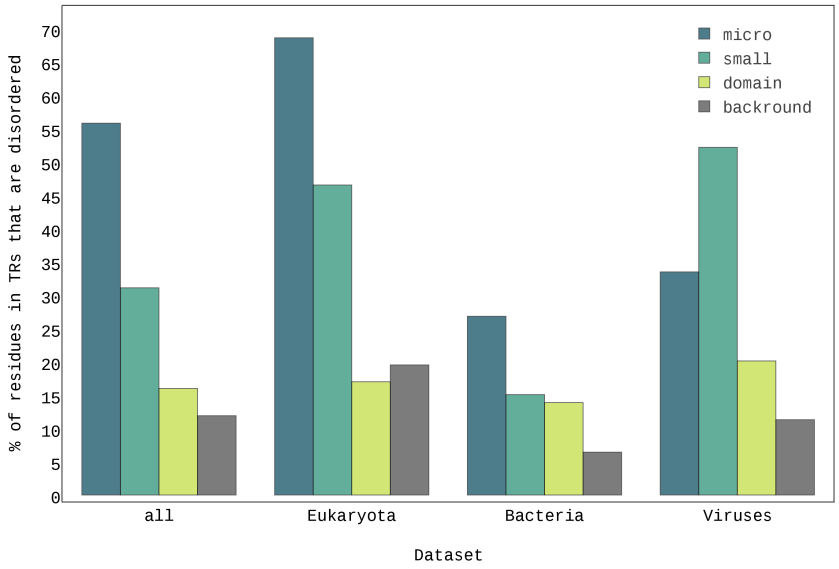


Figure 2:

[Figure X: % of disordered residues in TRs, split by unit size and kingdom]

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

TRs with small unit sizes are most disordered

[Figure X: % of disorder within a TR for different unit lengths]

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

[Figure X: AA frequencies in repeats vs background]

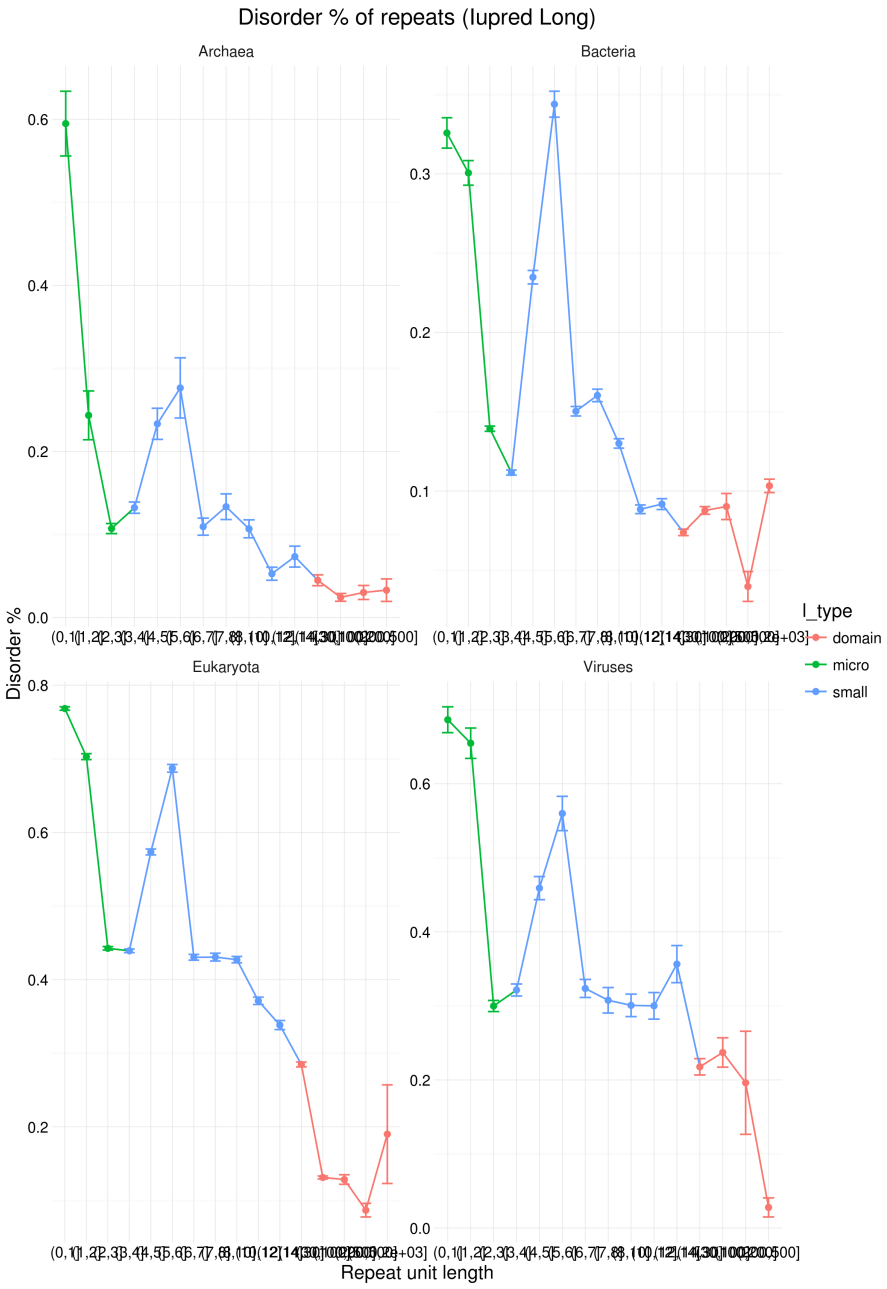

Disorder % of repeats (Iupred Long)
Archaea
Bacteria
Eukaryota
Viruses
Disorder %
Repeat unit length
l_type
domain
micro
small

Figure 3:

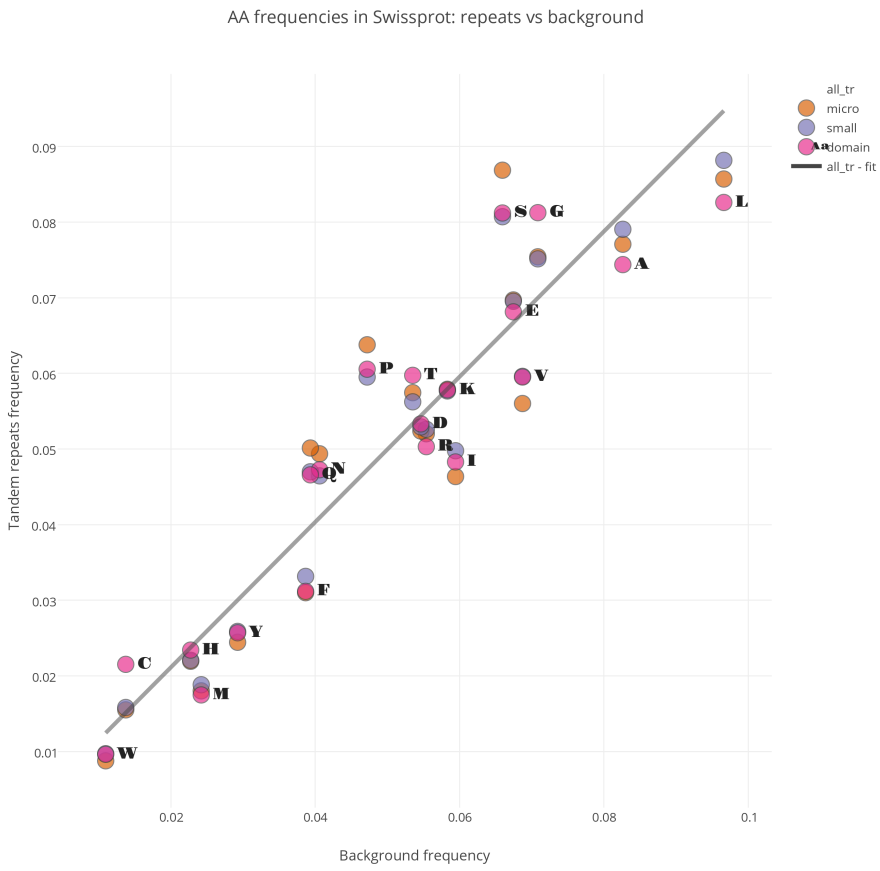

AA frequencies in Swissprot: repeats vs background
Tandem repeats frequency
Background frequency
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.02
0.04
0.06
0.08
0.1
all_tr
micro
small
domain
all_tr - fit
W
C
M
H
Y
F
N
Q
P
T
K
D
R
I
V
E
S
G
A
L

Figure 4:

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.]

Localization of TRs and disorder.

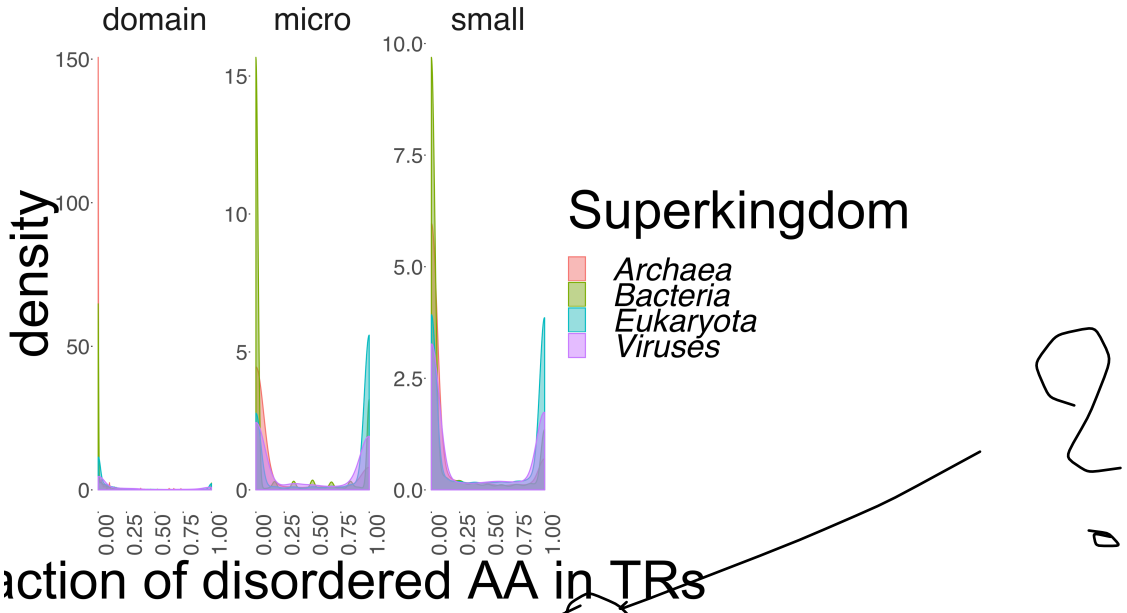


Fig. 4a) Shown is the fraction of disordered chars for micro TRs, small TRs and domain TRs. We see clearly bimodal distributions, with the majority TRs mostly ordered, however with a striking fraction of small TRs and micro TRs mostly disordered.

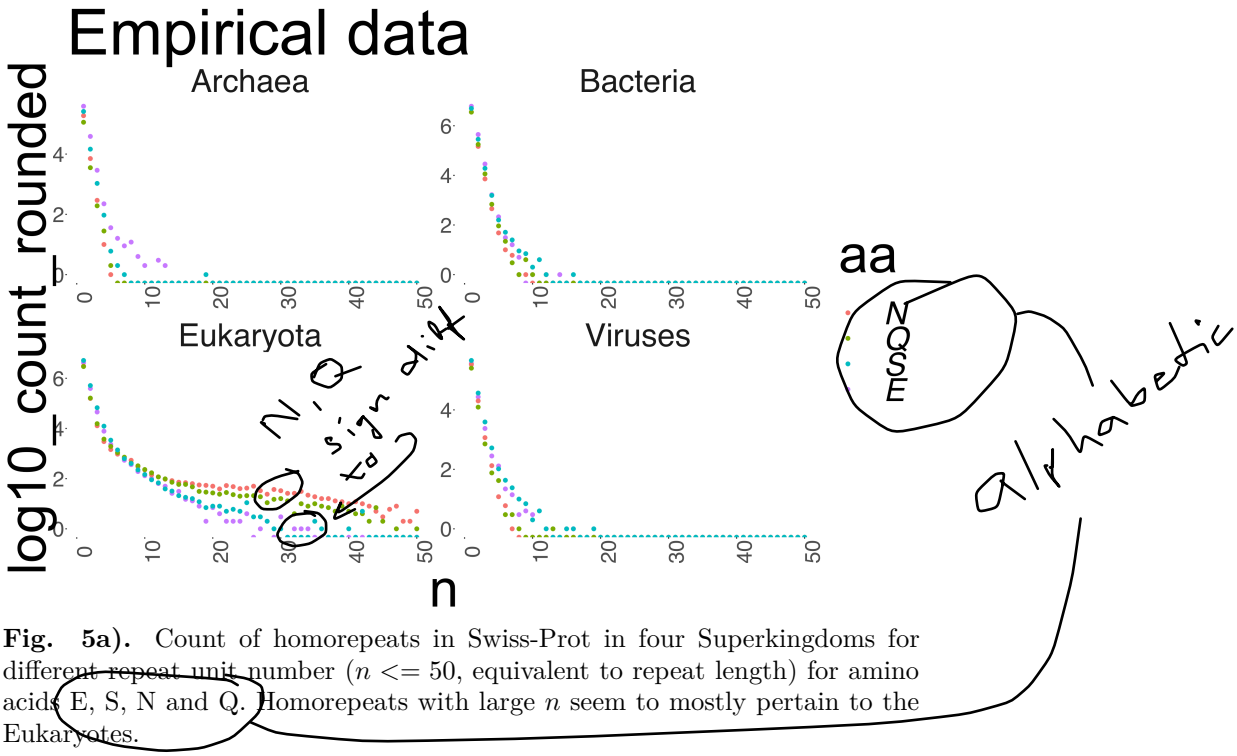


Fig. 5a). Count of homorepeats in Swiss-Prot in four Superkingdoms for different repeat unit number ($n <= 50$, equivalent to repeat length) for amino acids E, S, N and Q. Homorepeats with large n seem to mostly pertain to the Eukaryotes.

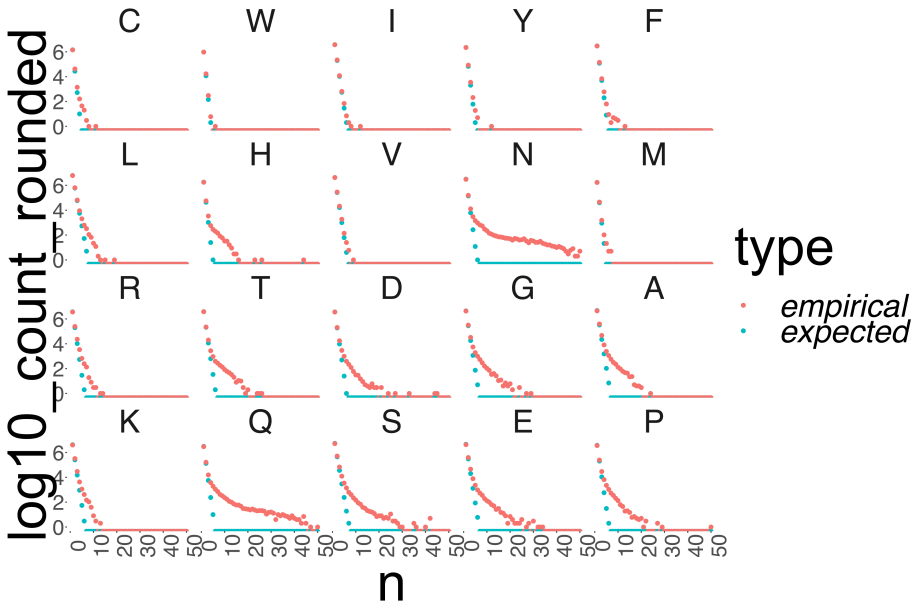


Fig. 5b). Empirical and expected count of homorepeats in Swiss-Prot Eukaryotes ($n <= 50$). Amino acids are ordered by their propensity to

promote structural order.

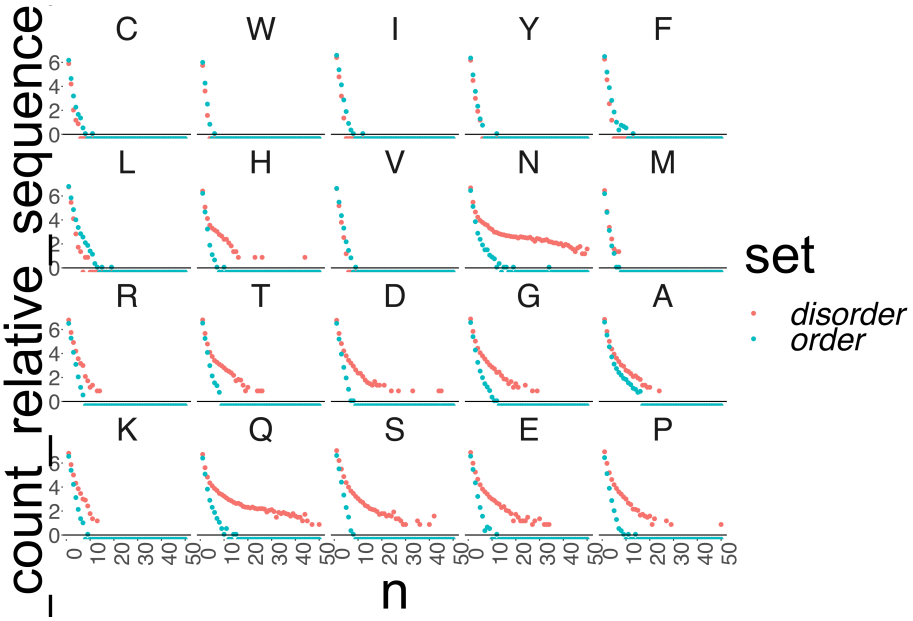


Fig. 5c). Empirical count of homorepeats in Swiss-Prot Eukaryotes ($n <= 50$) for ordered and disordered regions (consensus MobiDB annotations, no minimum length cut-off.). Amino acids are ordered by their propensity to promote structural order.

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

Conclusions

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.

newer data available?

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 protepries

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 short 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 {Schaper:22923522}.

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 {Schaper:24497029}. 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 {Schaper:24497029}, and subjected to stringent filtering (p-value=0.01).

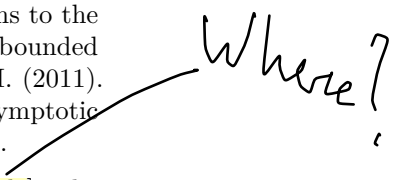
Homorepeat annotations

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..



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.