PLOS ONE

RESEARCH ARTICLE

Large Ankyrin repeat proteins are formed with similar and energetically favorable units

Ezequiel A. Galpern, María I. Freiberger, Diego U. Ferreiro*

Protein Physiology Lab, Departamento de Química Biológica, Instituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales (IQUIBICEN-CONICE), Universidad de Buenos Aires, Buenos Aires, Argentina

* ferreiro@qb.fcen.uba.ar

OPEN ACCESS

Citation: Galpern EA, Freiberger MI, Ferreiro DU (2020) Large Ankyrin repeat proteins are formed with similar and energetically favorable units. PLoS ONE 15(6): e0233865. https://doi.org/10.1371/journal.pone.0233865

Editor: Yaakov Koby Levy, Weizmann Institute of Science, ISRAEL

Received: February 4, 2020

Accepted: May 13, 2020

Published: June 24, 2020

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: DUF; Universidad de Buenos Aires-UBACYT 2018 - 20020170100540BA. http://www.uba.ar DUF; Agencia Nacional de Promoción Cienctífica y Tecnológica - PICT2016-1467 https://www.argentina.gob.ar/ciencia/agencia The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Ankyrin containing proteins are one of the most abundant repeat protein families present in all extant organisms. They are made with tandem copies of similar amino acid stretches that fold into elongated architectures. Here, we built and curated a dataset of 200 thousand proteins that contain 1.2 million Ankyrin regions and characterize the abundance, structure and energetics of the repetitive regions in natural proteins. We found that there is a continuous roughly exponential variety of array lengths with an exceptional frequency at 24 repeats. We described that individual repeats are seldom interrupted with long insertions and accept few deletions, in line with the known tertiary structures. We found that longer arrays are made up of repeats that are more similar to each other than shorter arrays, and display more favourable folding energy, hinting at their evolutionary origin. The array distributions show that there is a physical upper limit to the size of an array of repeats of about 120 copies, consistent with the limit found in nature. The identity patterns within the arrays suggest that they may have originated by sequential copies of more than one Ankyrin unit.

Introduction

Natural proteins that are formed with repetitions of stretches of amino-acids are abundant in extant organisms [1]. Some proteins contain repetitions of short stretches, forming fibrillate structures like collagen, and some contain longer repetitions of globular domains like beads on a string. In between, there is a class of proteins that is formed by tandem repetitions of similar stretches of about 30∼40 residues. These kinds of proteins (from now on repeat proteins) are present in all organisms and are believed to be ancient systems [2, 3]. Typically these polypeptides form elongated structures where each repeat motifs packs against its nearest neighbors, stabilizing an overall super-helical fold [4]. Since most of the structural characterization of these proteins were performed on model systems of short arrays that are experimentally amenable, we aim at characterizing the overall structures of an abundant family of proteins.

Ankyrin repeat proteins (ANKs) are usually described as formed with linear arrays of tandem copies of a 33 residues length motif that folds to a α-loop-α – β-hairpin/loop. Being one of the most common repeat proteins in nature, these molecules are believed to provide specific protein-protein interactions [5, 6]. Most of the structural knowledge about ANKs is derived

Competing interests: The authors have declared that no competing interests exist.

from the study of systems of biomedical relevance (the protein Ankyrin that gives name to the family, but also p16, Notch, IκB, etc, [7–10]); and from designed ANK proteins [11]. In these cases, the proteins are formed with a relatively few number of repeats, between 3 and 7, with a 12 repeat protein being the largest one for which folding was studied [12]. The folding of the repeat arrays can usually be described with a simple 1-D Ising model in which the most favourable repeats form a nuclei and structure propagates to near-neighbors [8, 12–15]. Small energetic inhomogeneities along the structure can break the folding cooperativity of multiple repeats and give rise to the appearance of folding intermediates [16–18]. Thus longer arrays are expected to break into folding subdomains of different stability [19, 20]. Moreover, good approximations to the folding energy can be constructed from statistical analysis of the extant sequences [21, 22]. We studied here the abundance, length distribution and energetics of ANK arrays in natural polypeptides.

In contrast to most globular domains, repeat proteins are believed to distinctively evolve by duplication and deletion of internal repetitions [2, 23–25]. It was recently suggested that the horizontal evolution is accelerated compared to their vertical divergence in related species [26]. The internal sequence similarity in each protein suggests that the repeats are often expanded through duplications of several repeats at a time, while the duplication of one repeat is less common, although no common mechanism for the expansion of repeats was found [25]. Here we re-examine the correlations of sequence similarity in ANKs and describe the occurrence of multiple types of duplication mechanisms within this family.

Methods

Repeats detection and array construction

In order to detect a majority of the possible Ankyrin repeats, we searched the full UniProtKB database [27], including manually reviewed Swiss-Prot (February 2019) and all the unreviewed TrEMBL (December 2017) sequences.

We used three structurally-derived hidden Markov models (HMM) developed by Parra et. al. [28] for ANK repeats: one HMM for internal repeats, one for C-terminal repeats and one for N-terminal repeats. These models fix a consistent phase for the repeat detection. We scanned all the database, splitted in single sequence fasta format, with the hmmsearch tool with default parameters [29] using the internal repeat HMM, detecting 194938 sequences with at least one hit. Subsequently, we ran hmmsearch with the other two HMM in order to detect terminal repeats and we eliminated the redundant hits.

To build an aligned repeats dataset from hmmer hits, we identified every model matched amino acids (AA) in the correspondent full protein sequence and copied AA before and after those detected that are needed to complete a 33 AA repeat. We took into account three particular cases: insertions inside the repeat, deletions and truncations. To resolve deletions and truncations, we simply admitted the gap character '-' in our AA alphabet. In the case of the insertions inside the repeat, we eliminated the corresponding positions for every insertion length. There is a possible case of double repeat detection, when hmmer identifies independently two hits which belong to the same repeat. After completing the repeats, we eliminated the double detection. We obtained a Multiple Repeat Alignment (MRA) of more than 1.2 million repeats sequences with exactly 33 positions.

In previous works, it has been reported that the insertions between ANK repeats have a characteristic length of 17 AA [28]. However, when they were analyzed at the full primary structure, we found a length distribution that extends beyond this (not shown). The distribution of insertions between repeats displayed a visible peak corresponding to a entire unit length of 33 AA. In these cases, we interpreted that the HMMs failed to detect a repeat between two

consecutive ones. Taking into account this observation, we defined an array as the concatenation of consecutive repeats that are less than 67 AA apart. With this definition, we considered the eventuality of missing a repeat in the HMM detection and an insertion of 17 AA each side the missed repeat. We note that we allowed to have more than one array for each full sequence, all of which we kept for analysis. Also, we note that the sequence database thus constructed does not necessarily represent the total universe of sequences, but is biased by the sequencing bias and by the phylogenetic relationship between the sequences. To minimize these biases in the analysis, we clustered the data by similarity using CD-hit [30] with a cutoff of 90% and we assigned a weight to the sequences defined as $1/n_i$, being n_i the number of sequences in the i th cluster. In this way, we end with 153209 effective arrays of ANK repeats. We took into account these weights to make all the statistical calculations.

Sequence identity calculations

We defined the pairwise identity or *pID* between two repeat sequences as the normalized quantity of identical AA in identical positions, excluding gap coincidences. We considered *pID* between every internal repeats in each array, distinguishing if they are first, second or i-th neighbors. N-terminal, C-terminal and internal repeats have been treated and characterized as different natural objects, with distinct sequence signatures [28]. In order to avoid border effects, we did not compare terminal to internal repeats in the pairwise identity analysis. Consistently, we considered arrays of four repeats onward, so each array has at least two internal repeats to compare to. We also considered the concept of sequence similarity when comparing repeats, using the BLOSUM62 matrix to score the alignments. Since we found that the main results are not altered by this analysis, we decided to use sequence identity as it is more sensitive to small sequence changes.

Autocorrelation analysis

We computed an auto correlation vector (*ACV*) between repeats r in an array as proposed by Björklund et al [25]. The n-component of the vector is the mean value of the *pID* for all r at neighborhood n, normalizing by the mean *pID* at first neighbors for the array

$$ACV_n = \frac{\langle pID_{(r_i,r_j)}\rangle_{|i-j|=n}}{\langle pID_{(r_i,r_j)}\rangle_{|i-j|=1}} \tag{1}$$

Energetic modeling

We considered that an Ankyrin repeat sequence is a state $\vec{\sigma} = (a_1, a_2, .., a_{L=33})$ as previously proposed [22]. Each position is occupied by one of the 20 amino-acids or the gap character, so it has 21 possibilities. We assumed that the system is in the state σ with a probability distribution that is mathematically equivalent to the Boltzmann distribution [22, 31]

$$P(\vec{\sigma}) = \frac{1}{Z} e^{-E(\vec{\sigma})} \tag{2}$$

taking the temperature such as $k_B\, T = 1$. Here $E(\vec{\sigma})$ is the energy of the state $\vec{\sigma}$ and Z is the partition function. If we assume that the positions are independent, discarding any interaction between different sites along the sequence, the energy can be written as

$$E(\vec{\sigma}) = \sum_i h_i(a_i). \tag{3}$$

where $h_i(a_i)$ is a local energy field that indicates the propensity to find an amino-acid a_i in a position i, and it can be calculated as follows using the frequency of finding in the MRA a residue in each column,

$$h_i(a_i) = -log[f_i(a_i)] + C. \quad (4)$$

We chose the constant C imposing the condition $\sum_{a_i} h_i(a_i) = 0$. The natural frequency $f_i(a_i)$ was measured taking into account the weights determined by the full sequence similarity clustering.

Results

Overall view of the dataset

The symmetrical nature of repeat-proteins allows the definition of units with a characteristic length of residues and a phase or initial position, that we identified and defined as *repeats*. In the case of Ankyrin proteins, considering preexisting structural studies [28], we worked with 33 amino-acids repeats and used the most common structural phase such that the TPLH motif occupies positions 10-13 in a repeat. However, repeats do not usually come alone in natural sequences, but one next to each other in long tandems conforming arrays. Given these definitions, we can find one or more array of repeats in each natural protein (Fig 1).

We collected and curated a database of 1.2 million repeats constructed as defined in *Methods* organized in 257703 arrays, which we weighted by phylogeny obtaining 153209 effective arrays. In 74% of the cases, all repeats in each protein cluster together in a single array, 19% of the proteins code for two arrays, 3% had three and only 4% have four or more arrays. Notably, there are example proteins that have up to 10 arrays. The effective arrays belong to Eukaryota

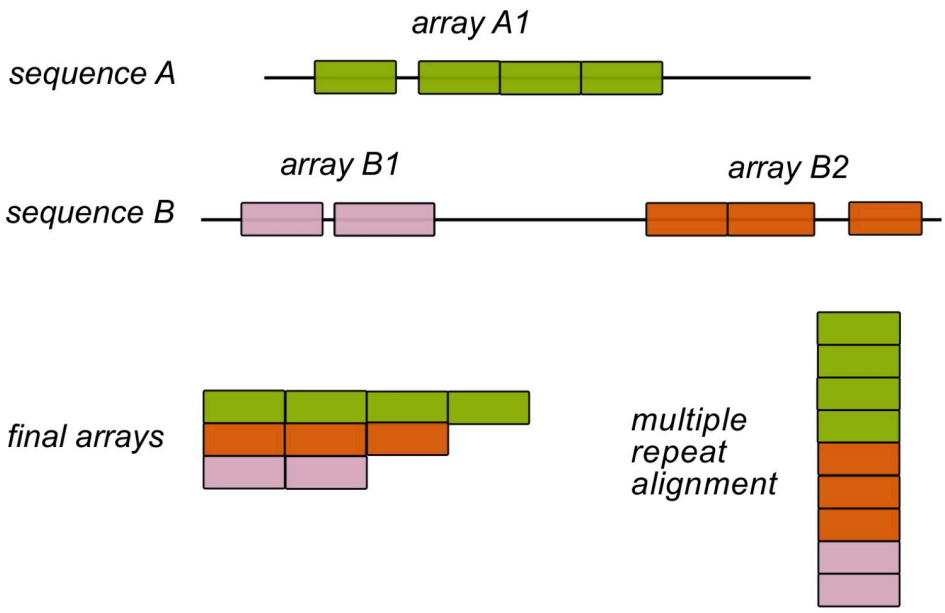

Fig 1. Definition of ankyrin arrays. We searched the whole UniProt database and detected repeats with a structurally-based HMM sequence model. If the detected repeats are separated by less than 67 residues, we defined them as belonging to the same array. In the above example, Sequence A codes for 1 single array, and sequence B codes for two arrays. Finally, we get a Multiple Repeat sequence Alignment (MRA) of more than 1.2 million repeats sequences with exactly 33 positions belonging to specific arrays.

https://doi.org/10.1371/journal.pone.0233865.g001

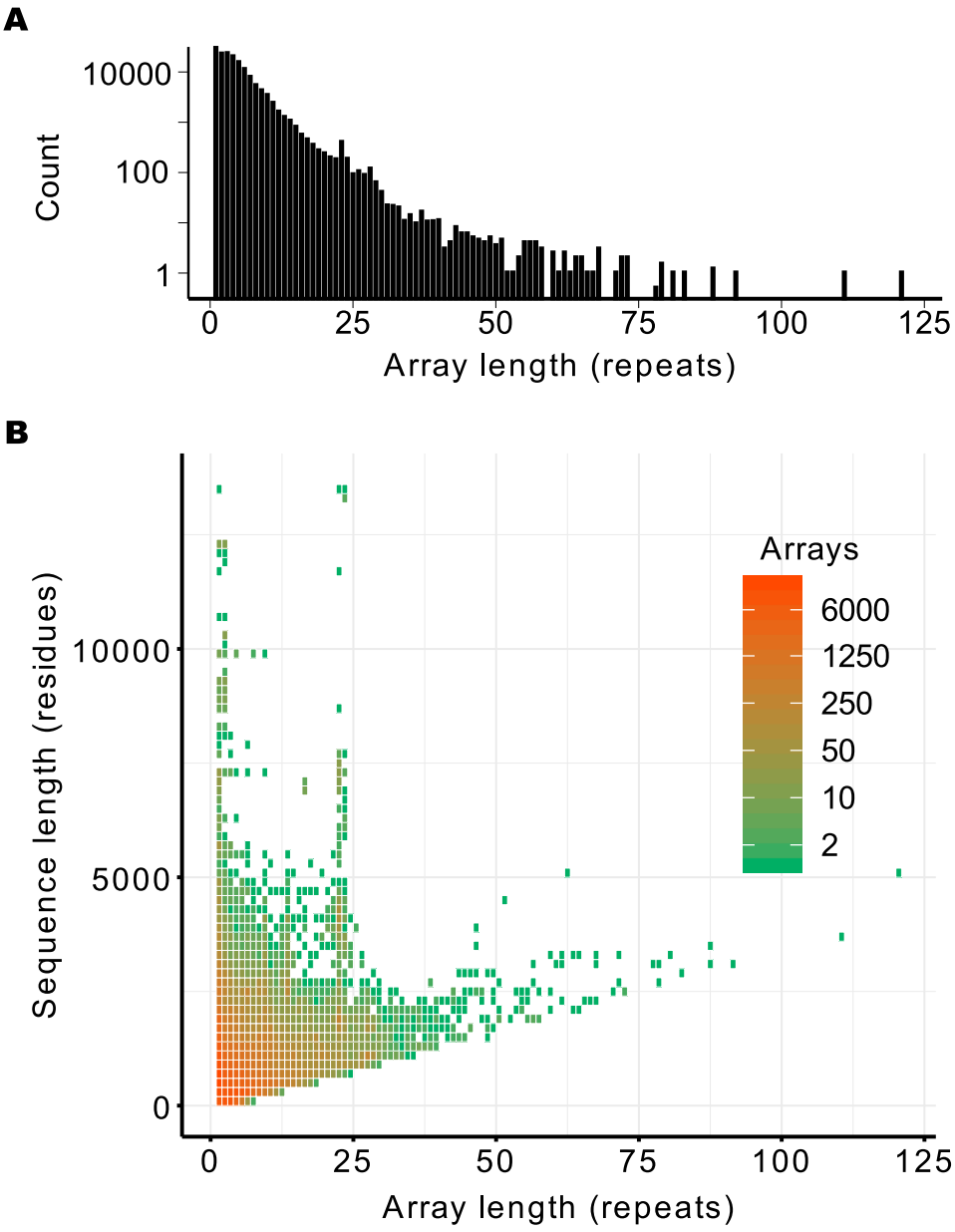

Fig 2. Array lengths in natural proteins. Natural sequences were collected and ANK arrays were constructed and weighted by phylogeny as described in *Methods*. A: The overall distribution of array lengths. B: array length compared to the total protein sequence length.

https://doi.org/10.1371/journal.pone.0233865.g002

proteomes in 85.5%, Bacteria 13.0%, Viruses 1.4% and Archaea 0.1%, in line with the previous census [1].

We classified the data according the array length, or simply the number of repeated units in each array. The distribution is presented in Fig 2A. There is a large number of arrays of just one repeat unit, representing 19% of arrays, of which 50% were detected as single repeats in the natural sequence and the remainder were at least 67 residues apart from their nearest neighbour. Since it is known that ANK proteins require multiple repeats to acquire a stable fold [15, 32, 33], these may represent miss detections of ANK patterns in unrelated sequences, as shown later by their energetic characterization (see below). The abundance of arrays

decreases roughly exponentially with array length with an anomalous peak around 23 repeats. The length distribution is not homogeneous across the domains of life, with the longest arrays being exclusively found in Eukarya (S1 Fig).

To analyze the distribution of arrays taking into account the total protein length, we combined the information in a heat map plot, presented in Fig 2B. There is a prohibited region in the plot, as sequences must have a minimum length of $33 * N$ to contain N units of 33 residues. The proteins for which all the polypeptide is formed with a single ANK array fall on the diagonal, notably up to one hundred repeats. On the upper left side of the plot there is a heterogeneity in the population distribution, with most proteins below 5000 residues containing arrays up to 25 repeat units. Still, there are examples of natural proteins over 10000 residues long that contain short arrays. Notably, the presence of arrays with 22∼23 repeats highlights in sequences from 3000 to 8000 residues long. It is interesting to note that there is one protein with an array of 23 repeats for which the crystallographic structure has been solved [34]. Analysis of this structure shows that our automatic repeat annotation missed one terminal repeat, and that the exact number of 24 ANK repeats corresponds with a complete turn of an ANK super-helix of $\sim 60\mathring{A}$ of diameter and $\sim 150\mathring{A}$ height [34]. Thus, the anomalous peak we detected in the length distribution 22∼24 may correspond to compact arrays of ANK repeats that make one complete turn when fully folded.

Natural ANK repeats do not always have exactly 33 residues [28]. Usually the structure can tolerate insertions, that we detected in the primary structure with the protocol described in *Methods*. We found that insertions occur only in 9% of the repeats of natural proteins. The distribution of the insertions length shows that the majority of them are of just one amino acid long, and insertions longer than 5 residues are rare (Fig 3A). The sites were the insertions occur along the ANK repeat are clearly not random (Fig 3B). Tertiary structure studies have previously characterized the insertion tolerance in ANK arrays [28] that is in excellent agreement with the ones we detected in the primary structure. There are two regions of the repeats where insertions are more likely, positions 6-7 and 17-20, that correspond with the linker regions between the helices that form the repeat units. Interestingly, we found repeats with long insertions of more than 60 AA in sequences of arrays between 3 and 10 repeats, reaching 1.2% of the repeats (S2 Fig). In some instances we found that a segment interpreted by us as an ANK repeat with a long insertion is annotated in Pfam as an ArfGAP domain, next to an ANK (e.g. Q9QWY8). In other cases, the segment is not annotated in Pfam or it is annotated belonging to the ANK clan. Thus, there are cases for which the ANK arrays can tolerate the insertion of a complete globular domain. Conversely, we found that deletions are very rare, present in only 1.4% of the repeats. In no case deletions exceed 14 amino acids per repeat and they are typically shorter, up to 3 residues (S3 Fig). We calculated whether deletions and insertions appear in specific repeats along the arrays. The distributions show that insertions and deletions are mostly homogeneously distributed along the arrays (S4 Fig). Surprisingly, we found that arrays of length 23 and 24 repeats present less deletions and insertions relative to the other ones. (S4 Fig). In summary, natural ANK arrays are tolerant to insertions in very specific positions in the repeats and are highly sensitive to deletions in their primary structure.

The longer the arrays, the more similar the repeats are

Are the ANK arrays constructed from a random sample of repeats or are there correlations between repeats that conform the arrays present in natural proteins? As a first step towards this analysis, we measured the pairwise identity at the sequence level (*pID*) between repeats, as described in *Methods*. We excluded from this analysis the terminal repeats of the arrays, and treated only internal repeats. Fig 4A shows the *pID* distribution for first neighbors repeats, that

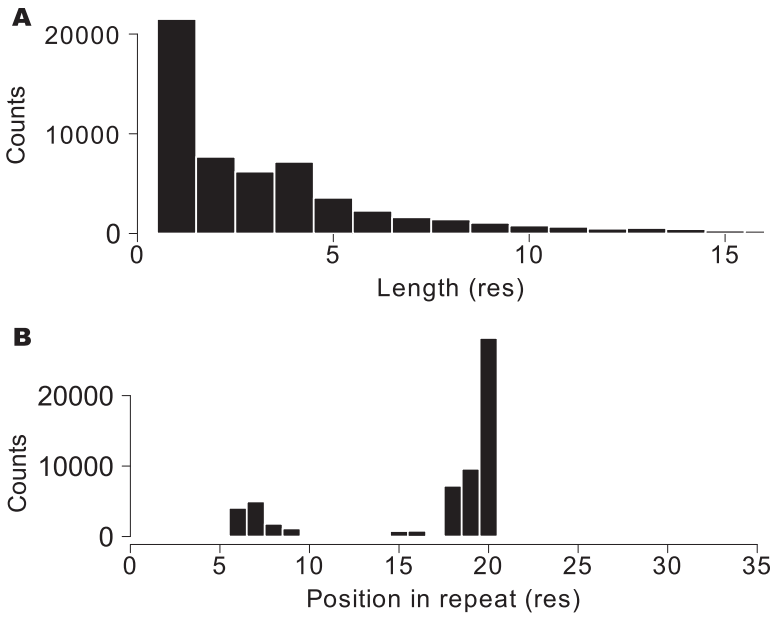

Fig 3. Intra-repeat insertions. A: Histogram of amino acid insertions lengths in ankyrin repeats. B: Histogram of relative position along the repeat where the insertions were found. The histogram indicates that insertions are more likely in positions 6-7 and 17-20.

https://doi.org/10.1371/journal.pone.0233865.g003

is to say consecutive repeats along the arrays, for arrays of different length. We compared the normalized distributions for arrays between 2 and 21 repeats long. The control group is an alignment of 2000 instances picked randomly from the entire alignment of internal repeats, keeping the proportions of each array length.

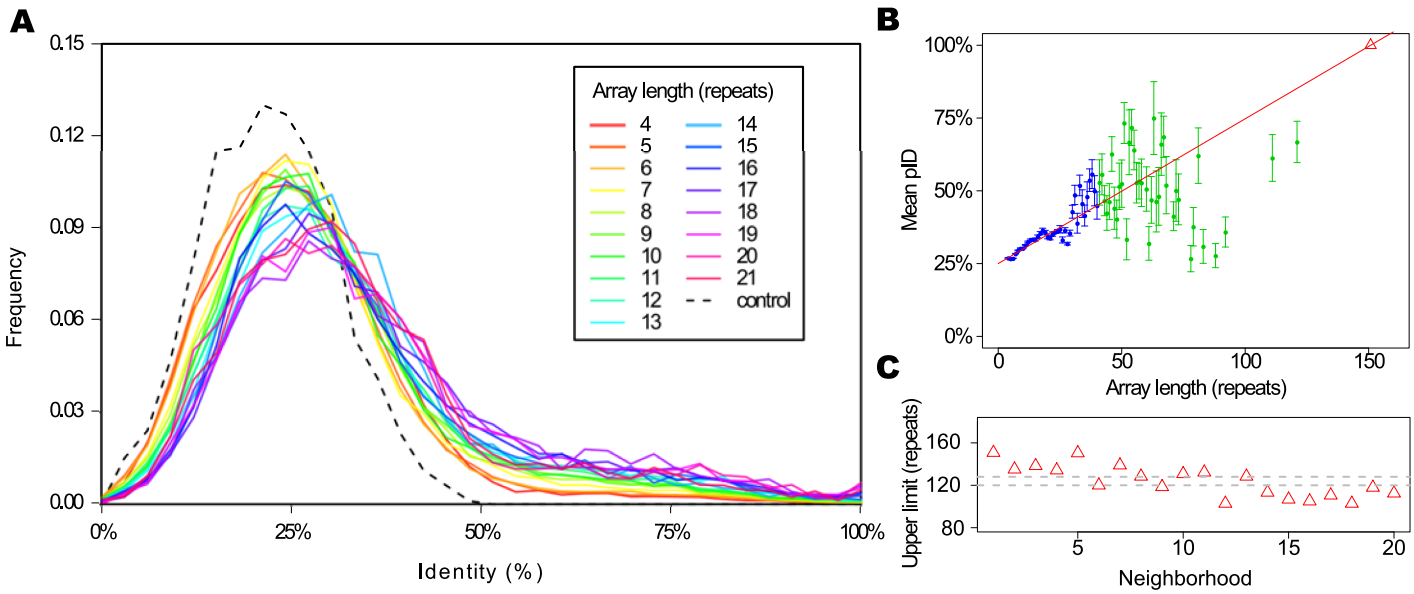

Fig 4. Pairwise identity. A: Pairwise identity at sequence level (*pID*) distribution between repeats sequence for first neighbors, for arrays of different length. We included results for arrays from 4 to 21 repeats in a rainbow color scale, normalized by the total counts. The control group is an alignment of 2000 repeats picked randomly from the entire alignment, holding the proportion of each array length. B: Mean values of identity between repeats sequence for first neighbors. The error bars indicates the standard error. The lineal fit (red line) only takes into account the blue points and errors, $R^2 = 0.9247$. Extrapolating for the maximum *pID* (100%) we find an upper limit for the array length (red triangle). C: Upper limit calculated for neighbors at different distances (red triangles) and the region defined by the mean and standard error of the points (grey dashed lines).

https://doi.org/10.1371/journal.pone.0233865.g004

We observe that all the natural distributions are distinguishable from the control one, that is to say, natural arrays are not constructed with random samples of repeats. The distributions peak below 25% for short arrays and shift smoothly as longer arrays are considered. The longer the arrays, the more similar the neighboring repeats are. The same trend appears when the *pID* of second neighboring repeats and onward is calculated (not shown). Fig 4B shows the mean *pID* for first neighbors, where we can clearly observe the mentioned trend, where repeats in longer arrays tend to be more similar between each other. This trend is better defined for the arrays up to 40 repeats, for which we have at least 10 different arrays in the database, and it get noisier for longer ones as the data gets sparser. Taking into account only the shorter arrays mean *pID* values and errors (blue points and bars in Fig 4B), we extrapolated linearly to an intercept with the line $mean(pID) = 100\%$ for an array length of 150 repeats. Repeating this analysis for neighbors at different distances the trend holds true (Fig 4C). By taking the mean and standard error we can define a region where we expect the upper limit of an array length composed of identical (124 ± 4) repeats (Fig 4C). This array length is coincident with the longer arrays found in the natural data set, and may constitute a physical upper limit for the length of an Ankyrin repeat array.

Correlations within the arrays

Are there consistent patterns in the distribution of repeats within the arrays? To investigate this question, we calculated the *pID* between all the pair-repeats in every array and analysed the resulting matrices. An example of such matrix is shown in Fig 5A. For this protein there is an evident chessboard pattern where repeats at distance of two neighbors appear to be more similar than consecutive ones. Also, the terminal repeats appear to be very different from the internal ones. A simple way to quantify this observation is to compute the autocorrelation vector (*ACV*) [25] with the *pID* as score, as detailed in *Methods*. In Fig 5B we present the corresponding *ACV* for this example protein, which has clear period of 2. Each component of the ACV is a mean value of a diagonal in the upper side of the matrix in Fig 5A, normalized by the mean *pID* at first neighbors for the array. It is important to notice that for neighbors farther apart the signal gets noisier merely because of the lack of data. The last element of the *ACV* vector is the normalized value for only one element of the *pID* matrix.

We made the *ACV* calculation for all the 257703 arrays, observing that many proteins have very different identifiable periods. There are proteins that present signals at lengths of 3, 5, 6 and 7 (S5, S6, S7 and S8 Figs), while other proteins present *ACV* with no appreciable signal. Also, we found examples of proteins that display two different periods along one single array (S6 Fig). We found that the distribution of patterns is not characteristic of single domains of life, but both Eukaryota and Bacteria encode proteins with various *ACV* distributions (S6 Fig). Another notable characteristic is the qualitative difference between the terminal repeats and the internal ones along the arrays, and in some cases between more than one terminal repeat and the rest of the array (Fig 5A for repeats 17 and 18).

In order to find if there is any general pattern for long proteins, we considered the arrays with 12 or more repeats and we calculate the *ACV* for each one up to neighborhood 7, only for internal repeats, and we then took the mean of all of them considering the phylogenetic biases as described in *Methods*. Using this subset of more than 11.4 thousand effective arrays allowed us to avoid the noisier components of each *ACV*. The overall signal is presented in Fig 6A and collects together a relative measurement of autocorrelation per array. The curve presents a maximum for neighborhood 2 and 4, where the relative identity is greater than that of the nearest neighbours. Also, the mean overall *ACV* decreases with the distance between repeats. For the same subset of arrays, we calculated the maximum for the *ACV* of each array (Fig 6B).

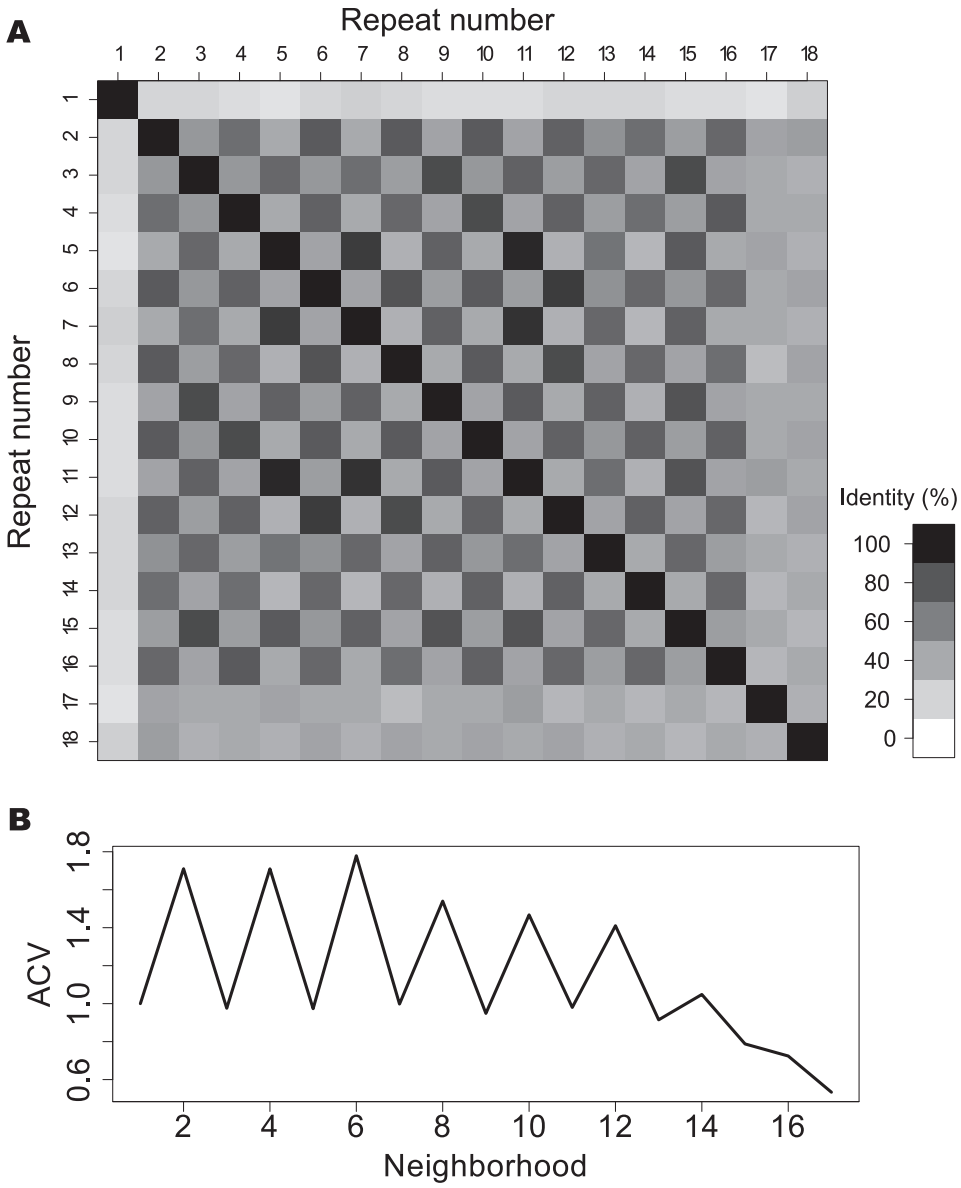

Fig 5. Autocorrelation vector of an ankyrin array. A: Pairwise identity matrix for the repeats of the W4XDH7 protein, that occupies positions 33-616 in the protein as an array of 18 repeats. B: Autocorrelation vector (*ACV*) for the same array.

https://doi.org/10.1371/journal.pone.0233865.g005

The nearest neighbors repeats have the greatest score in most cases. The distribution displays a weak decreasing trend, so the maximum of *ACV*s is sparse. The distribution of maximum *ACV* is roughly the same for Eukaryota and Bacteria (S9 Fig). Also, arrays with each maximum seem to be distributed without an evident trend along array length (S10 Fig). Finally, we calculated the mean *pID* per neighborhood for each array length (S11 Fig). On average, larger arrays present stronger periodicities than shorter ones, so the *ACV* signal that we obtained for every array is not a consequence of their overall similarity. In summary, the autocorrelation analysis of all the ANK repeat proteins points that the arrays are constructed with internal copies of various repeats, where sometimes the duplicated unit appears to be two repeats, sometimes three, five and up to seven consecutive units.

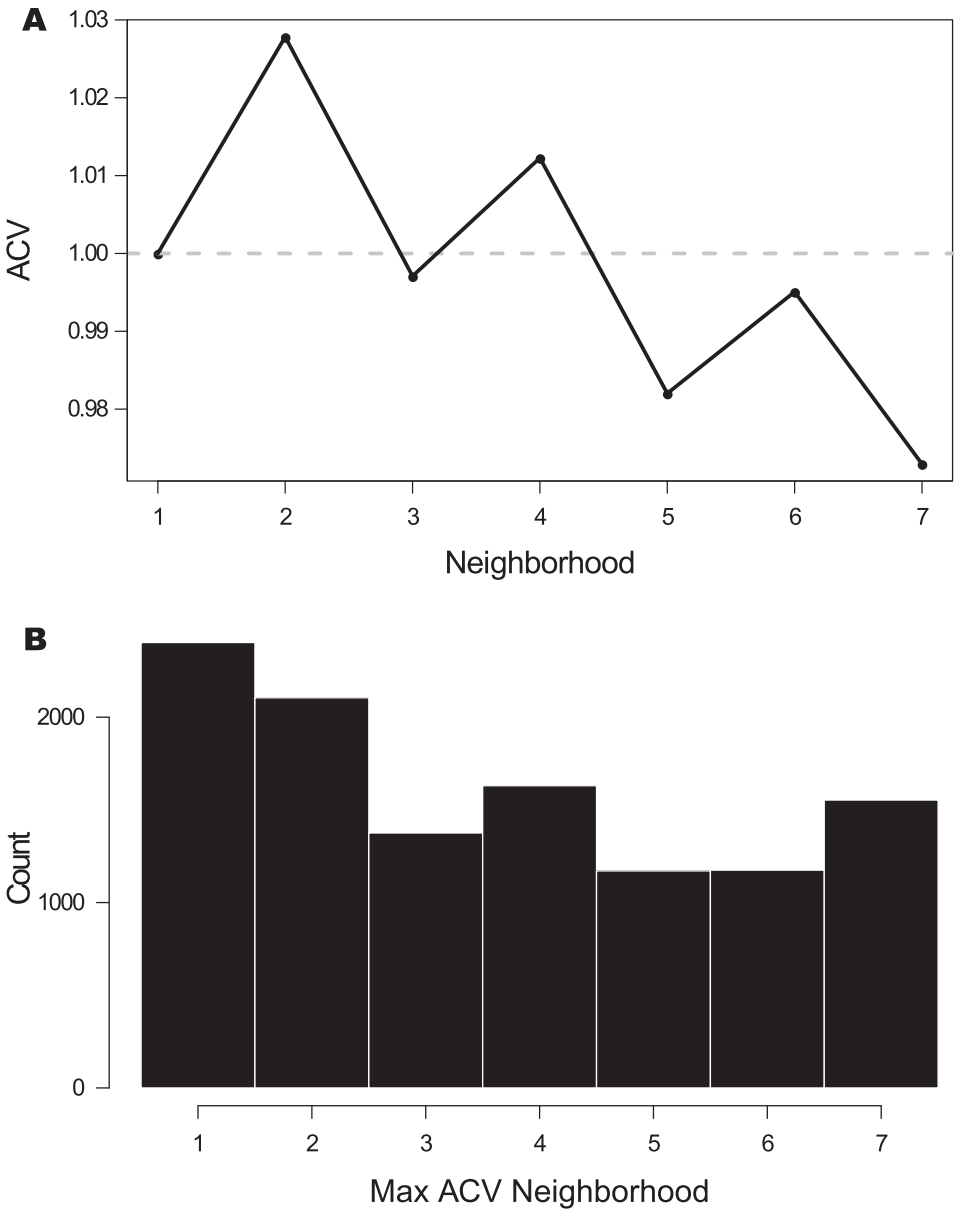

Fig 6. Average and maximum autocorrelation vector. A: Average autocorrelation vector (*ACV*) up to neighborhood 7 for arrays with 12 or more repeats, considering only internal repeats. The signal is normalized per array. B: Histogram of the maximum of each *ACV* for the same subset of arrays.

https://doi.org/10.1371/journal.pone.0233865.g006

Energetic characterization of the arrays

In order to analyze the folding energy distribution of the natural arrays found in protein sequences, we defined a simple energetic model based on the per-site occurrence of amino acids (see Methods). This model is a simplification of a previously reported one [21] that captures the most salient energetic features. We split up the Ankyrin repeat alignment into one alignment per array length and we calculated the energetic distribution for each case, taking into account the phylogenetic biases weighted as described in *Methods*. More negative energetic values indicate more favorable protein sequences. We show the energy

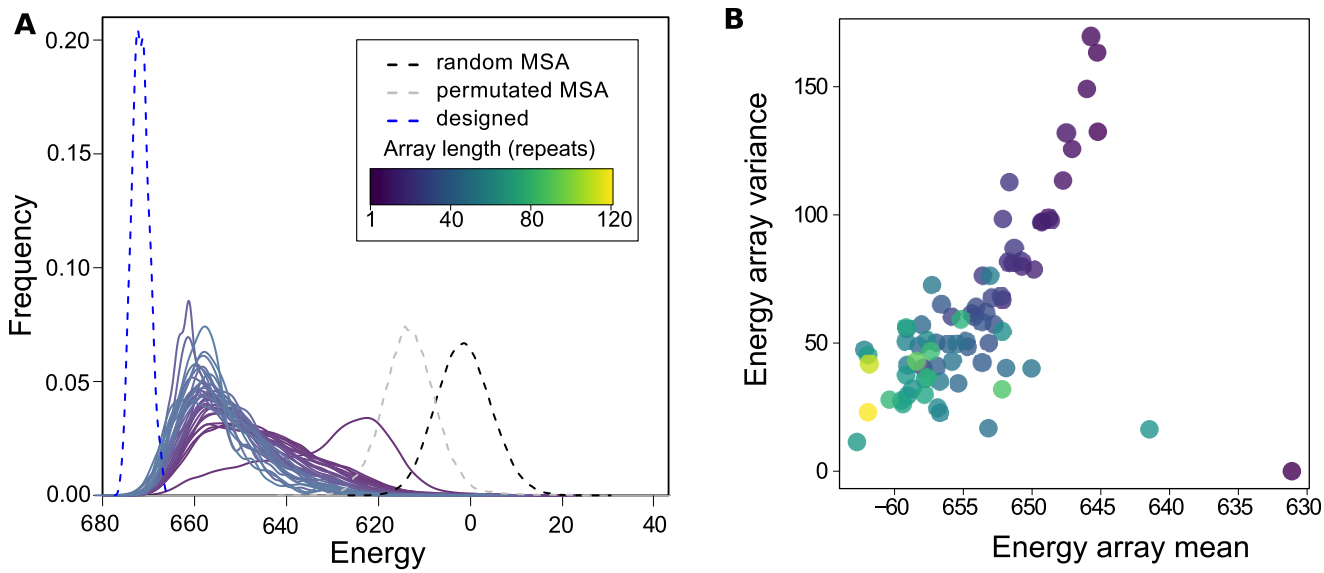

Fig 7. Energetic characterization. A: Energy distribution for repeats that belongs to arrays of different length, from 1 to 37 repeats. In black dashed line, the distribution for a random multiple repeat alignment (MRA). In dashed grey line, for the natural MRA with permuted columns and in dashed blue line the distribution for designed by consensus Ankyrin repeats. B: Energy mean and variance for each array, averaged according to the arrays length. The color scale is indicated in A.

https://doi.org/10.1371/journal.pone.0233865.g007

distributions for each array length and three references in Fig 7A. First, the energy distribution for an alignment of sequences of 33 random residues is centered near zero, as defined by the model. Second, the distribution for an alignment of consensus-like Ankyrin repeats [35], are clearly shifted to the lowest values, in correspondence to their measured extreme thermodynamic stability [36]. Finally, the energy distribution for the natural and complete alignment but with its columns permuted, thus keeping the natural amino acid distribution, fall in between the extremes. In Fig 7B we show the mean energy and variance for every array length.

The distribution that corresponds to repeats that come alone in the arrays is clearly distinguishable form the rest. The single repeats seems to be the least favorable in this energetic scale, and regarding the mean value the difference is higher (Fig 7). This indicates that single repeats collected in the database are different objects from the ones that come in pairs or bigger tandems and may even be considered as non-true Ankyrin repeats. Furthermore, single repeats that are alone in the full sequences or that share the protein with an other array have distributions with non-significant differences between them, indicating that the actual natural arrays are continuous tandem objects.

For repeats that come in pairs or longer tandems, the distributions clearly shift to more favourable regions as the arrays get longer, and the variance gets smaller. This observation is still evident when we eliminate from the analysis the terminal repeats, so it cannot be attributed to a border effect (not shown).

If we consider the energy of the consensus-designed proteins, the distribution is centered at -70 units, which appears to be the lower limit of the energy scale. In conclusion, longer arrays are formed with repeats that are more energetically favorable than repeats that form shorter arrays. Interestingly, it is clear that longer arrays are not only closer to the energy minimum, but are overall more homogeneous in their energy distribution (Fig 7B), indicating that they are formed with sequences that display similar local stabilization energy.

Discussion

We constructed a large dataset of Ankyrin-repeat arrays by collecting and curating sequences from all the known proteomes of a large variety of organisms. We analyzed one and a half hundred thousand non-redundant arrays containing more than 1.2 million aligned repeats. Around 75 percent of the proteins present a single array of multiple repeats. We found that 80 percent of the arrays are constituted with less than 7 repeats, yet the arrays span a large variety of sizes with roughly an exponential distribution (Fig 2). We found that insertions in the ANK repeats are rare, with both the length and the most common relative position of the insertions compatible with a previous 3D structural analysis [28] for the Ankyrin family. Curiously, we found a particularly abundant array length of 22-24 repeat-units. Structurally, this is the size needed for ANK arrays to make a complete turn of the superhelical fold [34], and thus may be exceptionally abundant for functional reasons, such as to bring in spatial proximity binding partners that are held together at each end of the repetitive array.

The analysis of the pairwise identity *pID* between repeats that belongs to arrays of different lengths shows that shorter arrays are less homogeneous, but longer ones impose, gradually, a higher *pID* between first neighbors. If extrapolated to conform an array of identical repeats, this trend implies an upper limit for the array length that we estimated to be (124 ± 4) repeats, which is compatible with the longest arrays found in natural proteins, although a detailed mechanistic justification for this apparent limit remains to be established. Considering a simple site-independent model to approximate the folding energy [21], we calculated the energy distributions of the arrays and found that longer arrays are made with more favorable repeats than shorter arrays (Fig 7A). At the same time, longer arrays are found to be more energetically homogeneous than shorter ones (Fig 7B). Energy landscape theory arguments [37] predict that non-native traps would raise bigger free energy barriers in the folding of large proteins, so selection against misfolding should be stronger for longer proteins than shorter ones. To avoid misfolded traps, repeat protein may have to be more homogeneous and favorable as they get longer, nucleating folding and propagating to near neighbours [8, 12–15, 38–40], which is in line with our findings in the natural proteins. We propose that long, heterogeneous and less favorable repeat arrays may not fold robustly *in vivo* and may be detrimental to fitness, so we will not find them in nature. Recently, Persi et al [26] proposed that there is a universal accelerated horizontal evolution of repeats that drive them to homogeneity, finding strong signatures of purifying selection, which is compatible with the scenario we propose.

Comparing the *pID* between the repeats of the same array at fix neighborhood using an autocorrelation vectors *ACV* analysis [25] reveals that there are, in many cases, clear periodicities along the tandem copies of the arrays. In some proteins, the array appears to be originated with copies of two consecutive Ankyrin repeats (Fig 5), while in other instances the pattern has periods from 3 to at least 7 repeats (S5, S6, S7 and S8 Figs), consistent with previous findings [25]. The size of our data set allowed us to get clear *ACV* signals, which averaged over the set *ACV* peaks in 2,4 and 6 repeats with a decreasing trend (Fig 6A). The distribution of absolute maxima for each protein is roughly uniform at least up to neighborhood 7 (Fig 6B). Björklund et al [25, 41] postulated that there may be a biological mechanism that can copy and insert more than one repeat at once, giving rise to *Superepeats* (SR) in the structure of repeat proteins. This could explain the uneven distribution of the *ACV*s, which is clear in the Nebulin family [41]. For ANKs, our results are compatible with the existence of SR with different lengths in particular cases. Given the roughly uniform distribution of maximum *ACV* (Fig 6B), we cannot point to a characteristic duplication size of the SR unit. This kind of expansion of internal repeats does not seem to have a characteristic length for the SR, but a weak decreasing probability as the number of repeat units by SR increases. However, if we look at particular

instances, proteins such as W4XDH7 (Fig 5) shows a regular periodicity in the *ACV*, indicating that the SR has copied several times in the same sequence and, notably, conserving the phase of the repeat unit. The same behaviour at other repeat frequencies is observed for W4ZBY3 (S5 Fig), for A0A0L8GA82 (S7 Fig) and for A0A1X7UVJ5 (S8 Fig).

Taken together, these results suggest that there may be a generative mechanism for duplicating units that depends on the identity of the existent repeats. Once a SR is copied, the next duplication event is biased in favor of the same SR length. In other words, the duplication mechanism should somehow recognize the previous SR copy as a seed to make a new copy. This "memory effect" of the last step could be explained with an identity dependent mechanism. We propose a molecular mechanism that at first copies any number of repeats at the same time and paste them in tandem with the preexisting ones. When this happens once, the probability of it happening again increases, preserving the phase and the number of copied units. However, we noted that there are also examples with two different periods along the same array, like the bacterial protein R5A1C8 (S6 Fig), which in this framework could indicate the generation of two independent "seeds" in the same sequence. The existence of harmonics in the copies explains why the average *ACV* is higher for second neighbors than for the first ones, even though there are more similar first neighbors than second ones.

Repeat duplication could be explained by various molecular mechanisms such as illegitimate recombination, exon shuffling, DNA slippage, etc., but no common mechanism for the expansion of all repeats could be detected [25]. We found that the distribution of maximum *ACV* is roughly the same in Eukaryota and Bacteria in the ANK family (S9 Fig).

It should be noted that even if a length-independent SR copying mechanism may be acting, physical folding limits prevent the existence of arbitrary long tandem ANK repeat-proteins. On the one hand, sequences can not be arbitrarily energetically favorable locally in each part of the array, as the internal and inter-repeat contacts are limited by the ANK topology. On the other hand, folding cracks are more likely to occur in long arrays, as the entropy gain of introducing a defect grows with chain length [14].

Supporting information

S1 Fig. Array length according to cell type. Histogram of array length for Eukaryota (red), Bacteria (blue), Viruses (green) and Archaea (orange).
(EPS)

S2 Fig. Insertion length and repeat number per sequence heat map.
(EPS)

S3 Fig. Histogram of deletions length.
(EPS)

S4 Fig. Heat maps of insertion and deletion density per repeat for each array length and for each unit position. A: Density of insertions presence per repeat. B: Density of deletions presence per repeat. C: Density of inserted residues per repeat. D: Density of deleted positions per repeat.
(EPS)

S5 Fig. Autocorrelation vector for W4ZBY3. A: Pairwise identity matrix for the W4ZBY3 protein, positions 31-2394, an array of 71 repeats. B: Autocorrelation vector (*ACV*) for the same array.
(EPS)

S6 Fig. Autocorrelation vector for R5A1C8. A: Pairwise identity matrix for the bacterial protein R5A1C8, positions 1-651, an array of 20 repeats. B: Autocorrelation vector (*ACV*) for the same array.
(EPS)

S7 Fig. Autocorrelation vector for A0A0L8GA82. A: Pairwise identity matrix for the A0A0L8GA82 protein, positions 9-545, an array of 16 repeats. B: Autocorrelation vector (*ACV*) for the same array.
(EPS)

S8 Fig. Autocorrelation vector for A0A1X7UVJ5. A: Pairwise identity matrix for the A0A1X7UVJ5 protein, positions 674-2154, an array of 38 repeats. B: Autocorrelation vector (*ACV*) for the same array.
(EPS)

S9 Fig. Maximum autocorrelation vector according to cell type. Histograms for Eukaryota (red) and Bacteria (blue) of the maximum of each Autocorrelation vector (*ACV*) up to neighborhood 7 for arrays with 12 or more repeats, only for internal repeats.
(EPS)

S10 Fig. Maximum autocorrelation vector according to array length. Histograms for different array length of the maximum of each Autocorrelation vector (*ACV*) up to neighborhood 7 for arrays with 12 or more repeats, only for internal repeats.
(EPS)

S11 Fig. Mean pID per neighborhood for each array length, only for internal repeats.
(EPS)

Acknowledgments

We thank R. Gonzalo Parra, Rocío Espada, Juliana Glavina, Ariel Aptekmann and Hugo Salas for their fruitful comments and the Protein Physiology Lab team for their support.

Author Contributions

Conceptualization: Ezequiel A. Galpern, Diego U. Ferreiro.

Data curation: Ezequiel A. Galpern, María I. Freiberger.

Formal analysis: Ezequiel A. Galpern, Diego U. Ferreiro.

Funding acquisition: Diego U. Ferreiro.

Investigation: Ezequiel A. Galpern, María I. Freiberger.

Methodology: Ezequiel A. Galpern.

Project administration: Diego U. Ferreiro.

Resources: Diego U. Ferreiro.

Software: María I. Freiberger.

Supervision: Diego U. Ferreiro.

Validation: Diego U. Ferreiro.

Visualization: Ezequiel A. Galpern.

Writing – original draft: Ezequiel A. Galpern.

Writing – review & editing: Ezequiel A. Galpern, María I. Freiberger, Diego U. Ferreiro.

References

1. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. A census of protein repeats. J Mol Biol. 1999; 293 (1):151–60. PMID: 10512723
2. Schüler A, Bornberg-Bauer E. Evolution of Protein Domain Repeats in Metazoa. Mol Biol Evol. 2016; 33 (12):3170–3182. https://doi.org/10.1093/molbev/msw194 PMID: 27671125
3. Reshef D, Itzhaki Z, Schueler-Furman O. Increased sequence conservation of domain repeats in prokaryotic proteins. Trends in genetics. 2010; 26(9):383–387. PMID: 20643491
4. Paladin L, Hirsh L, Piovesan D, Andrade-Navarro MA, Kajava AV, Tosatto SCE. RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures. Nucleic Acids Res. 2017; 45(D1):D308–D312. https://doi.org/10.1093/nar/gkw1136 PMID: 27899671
5. Li J, Mahajan A, Tsai MD. Ankyrin repeat: a unique motif mediating protein-protein interactions. Biochemistry. 2006; 45(51):15168–78. PMID: 17176038
6. Jernigan KK, Bordenstein SR. Ankyrin domains across the Tree of Life. PeerJ. 2014; 2:e264. https://doi.org/10.7717/peerj.264 PMID: 24688847
7. Tang KS, Guralnick BJ, Wang WK, Fersht AR, Itzhaki LS. Stability and folding of the tumour suppressor protein p16. J Mol Biol. 1999; 285(4):1869–86. PMID: 9917418
8. Street TO, Barrick D. Predicting repeat protein folding kinetics from an experimentally determined folding energy landscape. Protein Sci. 2009; 18(1):58–68. https://doi.org/10.1002/pro.9 PMID: 19177351
9. DeVries I, Ferreiro DU, Sánchez IE, Komives EA. Folding kinetics of the cooperatively folded subdomain of the *IκBα* ankyrin repeat domain. J Mol Biol. 2011; 408(1):163–76.
10. Barrick D, Ferreiro DU, Komives EA. Folding landscapes of ankyrin repeat proteins: experiments meet theory. Curr Opin Struct Biol. 2008; 18(1):27–34. https://doi.org/10.1016/j.sbi.2007.12.004 PMID: 18243686
11. Tamaskovic R, Simon M, Stefan N, Schwill M, Plückthun A. Designed ankyrin repeat proteins (DARPins) from research to therapy. Methods Enzymol. 2012; 503:101–34. PMID: 22230567
12. Werbeck ND, Rowling PJE, Chellamuthu VR, Itzhaki LS. Shifting transition states in the unfolding of a large ankyrin repeat protein. Proc Natl Acad Sci U S A. 2008; 105(29):9982–7. https://doi.org/10.1073/pnas.0705300105 PMID: 18632570
13. Ferreiro DU, Wolynes PG. The capillarity picture and the kinetics of one-dimensional protein folding. Proc Natl Acad Sci U S A. 2008; 105(29):9853–4. https://doi.org/10.1073/pnas.0805287105 PMID: 18632565
14. Ferreiro DU, Walczak AM, Komives EA, Wolynes PG. The energy landscapes of repeat-containing proteins: topology, cooperativity, and the folding funnels of one-dimensional architectures. PLoS Comput Biol. 2008; 4(5):e1000070. https://doi.org/10.1371/journal.pcbi.1000070 PMID: 18483553
15. Aksel T, Barrick D. Direct observation of parallel folding pathways revealed using a symmetric repeat protein system. Biophys J. 2014; 107(1):220–32. https://doi.org/10.1016/j.bpj.2014.04.058 PMID: 24988356
16. Hutton RD, Wilkinson J, Faccin M, Sivertsson EM, Pelizzola A, Lowe AR, et al. Mapping the Topography of a Protein Energy Landscape. J Am Chem Soc. 2015; 137(46):14610–25. PMID: 26561984
17. Street TO, Bradley CM, Barrick D. Predicting coupling limits from an experimentally determined energy landscape. Proc Natl Acad Sci U S A. 2007; 104(12):4907–12. https://doi.org/10.1073/pnas.0608756104 PMID: 17360387
18. Ferreiro DU, Cervantes CF, Truhlar SME, Cho SS, Wolynes PG, Komives EA. Stabilizing IkappaBalpha by "consensus" design. J Mol Biol. 2007; 365(4):1201–16. https://doi.org/10.1016/j.jmb.2006.11.044
19. Michaely P, Bennett V. The membrane-binding domain of ankyrin contains four independently folded subdomains, each comprised of six ankyrin repeats. J Biol Chem. 1993; 268(30):22703–9. PMID: 8226780
20. Espada R, Parra RG, Sippl MJ, Mora T, Walczak AM, Ferreiro DU. Repeat proteins challenge the concept of structural domains. Biochem Soc Trans. 2015; 43(5):844–9. PMID: 26517892
21. Espada R, Parra RG, Mora T, Walczak AM, Ferreiro DU. Inferring repeat-protein energetics from evolutionary information. PLoS Comput Biol. 2017; 13(6):e1005584. https://doi.org/10.1371/journal.pcbi.1005584 PMID: 28617812

22. Marchi J, Galpern EA, Espada R, Ferreiro DU, Walczak AM, Mora T. Size and structure of the sequence space of repeat proteins. PLoS Comput Biol. 2019; 15(8):e1007282. https://doi.org/10.1371/journal.pcbi.1007282 PMID: 31415557

23. Andrade MA, Perez-Iratxeta C, Ponting CP. Protein repeats: structures, functions, and evolution. J Struct Biol. 2001; 134(2-3):117–31. PMID: 11551174

24. Pâques F, Leung WY, Haber JE. Expansions and contractions in a tandem repeat induced by double-strand break repair. Mol Cell Biol. 1998; 18(4):2045–54. https://doi.org/10.1128/mcb.18.4.2045 PMID: 9528777

25. Björklund ÅK, Ekman D, Elofsson A. Expansion of protein domain repeats. PLoS Computational Biology. 2006; 2(8):0959–0970.

26. Persi E, Wolf YI, Koonin EV. Positive and strongly relaxed purifying selection drive the evolution of repeats in proteins. Nature communications. 2016; 7:13570. https://doi.org/10.1038/ncomms13570 PMID: 27857066

27. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. Methods Mol Biol. 2007; 406:89–112. PMID: 18287689

28. Parra RG, Espada R, Verstraete N, Ferreiro DU. Structural and Energetic Characterization of the Ankyrin Repeat Protein Family. PLoS Computational Biology. 2015; 11(12):1–20.

29. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011; 7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195 PMID: 22039361

30. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22(13):1658–9. PMID: 16731699

31. Mora T, Bialek W. Are biological systems poised at criticality? Journal of Statistical Physics. 2011; 144 (2):268–302.

32. Rowling PJE, Sivertsson EM, Perez-Riba A, Main ERG, Itzhaki LS. Dissecting and reprogramming the folding and assembly of tandem-repeat proteins. Biochem Soc Trans. 2015; 43(5):881–8. PMID: 26517898

33. Ferreiro DU, Cho SS, Komives EA, Wolynes PG. The energy landscape of modular repeat proteins: topology determines folding mechanism in the ankyrin family. J Mol Biol. 2005; 354(3):679–92. PMID: 16257414

34. Wang C, Wei Z, Chen K, Ye F, Yu C, Bennett V, et al. Structural basis of diverse membrane target recognitions by ankyrins. eLife. 2014; 3(November):1–22.

35. Kohl A, Binz HK, Forrer P, Stumpp MT, Plückthun A, Grütter MG. Designed to be stable: crystal structure of a consensus ankyrin repeat protein. Proc Natl Acad Sci U S A. 2003; 100(4):1700–5. https://doi.org/10.1073/pnas.0337680100 PMID: 12566564

36. Binz HK, Stumpp MT, Forrer P, Amstutz P, Plückthun A. Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. J Mol Biol. 2003; 332(2):489–503. PMID: 12948497

37. Wolynes PG. Evolution, energy landscapes and the paradoxes of protein folding. Biochimie. 2015; 119:218–230. https://doi.org/10.1016/j.biochi.2014.12.007 PMID: 25530262

38. Wright CF, Teichmann SA, Clarke J, Dobson CM. The importance of sequence diversity in the aggregation and evolution of proteins. Nature. 2005; 438(7069):878–881. PMID: 16341018

39. Aksel T, Majumdar A, Barrick D. The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. Structure. 2011; 19(3):349–360. https://doi.org/10.1016/j.str.2010.12.018 PMID: 21397186

40. Hagai T, Azia A, Trizac E, Levy Y. Modulation of folding kinetics of repeat proteins: interplay between intra-and interdomain interactions. Biophysical journal. 2012; 103(7):1555–1565. https://doi.org/10.1016/j.bpj.2012.08.018 PMID: 23062348

41. Åsa K Björklund, Light S, Sagit R, Elofsson A. Nebulin: A Study of Protein Repeat Evolution. Journal of Molecular Biology. 2010; 402(1):38–51.