HAL
open science

Size and structure of the sequence space of repeat proteins

Jacopo Marchi, Ezequiel A Galpern, Rocío Espada, Diego U Ferreiro, Aleksandra M. Walczak, Thierry Mora

To cite this version:

Jacopo Marchi, Ezequiel A Galpern, Rocío Espada, Diego U Ferreiro, Aleksandra M. Walczak, et al.. Size and structure of the sequence space of repeat proteins. PLoS Computational Biology, 2019, 15 (8), pp.e1007282. 10.1371/journal.pcbi.1007282 . hal-02345012

HAL Id: hal-02345012

https://hal.sorbonne-universite.fr/hal-02345012v1

Submitted on 4 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Size and structure of the sequence space of repeat proteins

Jacopo Marchi[1], Ezequiel A. Galpern[2,3], Rocio Espada[4], Diego U. Ferreiro[2,3], Aleksandra M. Walczak[1☯*], Thierry Mora[1☯*]

1 Laboratoire de physique de l'École normale supérieure (PSL University), CNRS, Sorbonne Université, and Université de Paris, 75005 Paris, France, **2** Protein Physiology Lab, Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Química Biológica, Buenos Aires, Argentina, **3** CONICET - Universidad de Buenos Aires, Instituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales (IQUIBICEN), Buenos Aires, Argentina, **4** Laboratoire Gulliver, Ecole supérieure de physique et chimie industrielles (PSL University) and CNRS, 75005, Paris, France

☯ These authors contributed equally to this work.
* awalczak@lpt.ens.fr (AMW); tmora@lps.ens.fr (TM)

Check for updates

OPEN ACCESS

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

Editor: Avner Schlessinger, Icahn School of Medicine at Mount Sinai, UNITED STATES

Received: May 11, 2019

Accepted: July 24, 2019

Published: August 15, 2019

Copyright: © 2019 Marchi et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

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

Funding: This work was partly supported by European Research Council (erc.europa.eu) Consolidator Grant no 724208 to AMW and ECOS Sud – CONICYT (www.univ-paris13.fr/cofecub-ecos/ecos-sud) grant no. A14E04 to DUF and AMW. NASA Astrobiology Institute, Grant no. 80NSSC18M0093. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The coding space of protein sequences is shaped by evolutionary constraints set by requirements of function and stability. We show that the coding space of a given protein family—the total number of sequences in that family—can be estimated using models of maximum entropy trained on multiple sequence alignments of naturally occuring amino acid sequences. We analyzed and calculated the size of three abundant repeat proteins families, whose members are large proteins made of many repetitions of conserved portions of ~ 30 amino acids. While amino acid conservation at each position of the alignment explains most of the reduction of diversity relative to completely random sequences, we found that correlations between amino acid usage at different positions significantly impact that diversity. We quantified the impact of different types of correlations, functional and evolutionary, on sequence diversity. Analysis of the detailed structure of the coding space of the families revealed a rugged landscape, with many local energy minima of varying sizes with a hierarchical structure, reminiscent of fustrated energy landscapes of spin glass in physics. This clustered structure indicates a multiplicity of subtypes within each family, and suggests new strategies for protein design.

Author summary

Natural protein molecules are only a small subset of the possible strings of amino acids. This naturally calls the question of how many protein sequences theoretically exist that are functional, and how many have already been explored by nature. To help answer this question, we developed a statistical method to calculate the total potential number of protein sequences of a given family, focusing on three families of repeat proteins, which play important roles in a variety of cellular processes. The number of sequences that we compute is limited by functional interactions between the residues of the protein, as well as its evolutionary history. Applying techniques from the physics of disordered systems, we

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

show that the space of sequences has a rugged structure, which could hinder their evolution. Individual proteins can be organised into distinct clusters corresponding to basins of attraction of the landscape, suggesting the existence of subfamilies within each family.

Introduction

Natural proteins contain a record of their evolutionary history, as selective pressure constrains their amino-acid sequences to perform certain functions. However, if we take all proteins found in nature, their sequence appears to be random, without any apparent rules that distinguish their sequences from arbitrary polypeptides. Nonetheless, the volume of sequence space taken up by existing proteins is very small compared to all possible polypeptide strings of a given length [1], even more so when specializing to a given structure [2]. Clearly, not all variants are equally likely to survive [3–5]. To better understand the structure of the space of natural proteins, it is useful to group them into families of proteins with similar fold, function, and sequence, believed to be under a common selective pressure. Assuming that the ensemble of protein families is equilibrated, there should exist a relationship between the conserved features of their amino acid sequences and their function. This relation can be extracted by examining statistics of amino-acid composition, starting with single sites in multiple alignments (as provided by e.g. PFAM [6, 7]). More interesting information can be extracted from covariation of amino acid usages at pairs of positions [8–10] or using machine-learning techniques [11]. Models of protein sequences based of pairwise covariations have been shown to successfully predict pair-wise amino-acid contacts in three dimensional structures [12–17], aid protein folding algorithms [18, 19], and predict the effect of point mutations [17, 20–22]. However, little is known on how these identified amino-acid constraints affect the global size, shape and structure of the sequence space. Accounting for these questions is a first step towards drawing out the possible and the realized evolutionary trajectories of protein sequences [23, 24].

We use tools and concepts from the statistical mechanics of disordered systems to study collective, protein-wide effects and to understand how evolutionary constraints shape the landscape of protein families. We go beyond previous work which focused on local effects—pairwise contacts between residues, effect of single amino-acid mutations—to ask how amino-acid conservation and covariation restrict and shape the landscape of sequences in a family. Specifically, we characterize the size of the ensemble, defined as the effective number of sequences of a familiy, as well as its detailed structure: is it made of one block or divided into clusters of "basins"? These are intrinsically collective properties that can not be assessed locally.

Repeat proteins are excellent systems in which to quantify these collective effects, as they combine both local and global interactions. Repeat proteins are found as domains or subdomains in a very large number of functionally important proteins, in particular signaling proteins (e.g. NF-κB, p16, Notch [25]). Usually they are composed of tandem repetitions of ~ 30 amino-acids that fold into elongated architectures. Repeat proteins have been divided into different families based on their structural similarity. Here we consider three abundant repeat protein families: ankyrin repeats (ANK), tetratricopeptide repeats (TPR), leucine-rich repeat (LRR) that fold into repetitive structures (see Fig 1). In addition to interactions between residues within one repeat, repeat protein evolution is constrained by inter-repeat interactions, which lead to the characteristic accordeon-like folds. Through these separable types of constraints, as well as the possibility of intra- and inter-familly comparisons, repeat proteins are perfect candidates to ask questions about the origins and the effects of the constraints that globally shape the sequences.

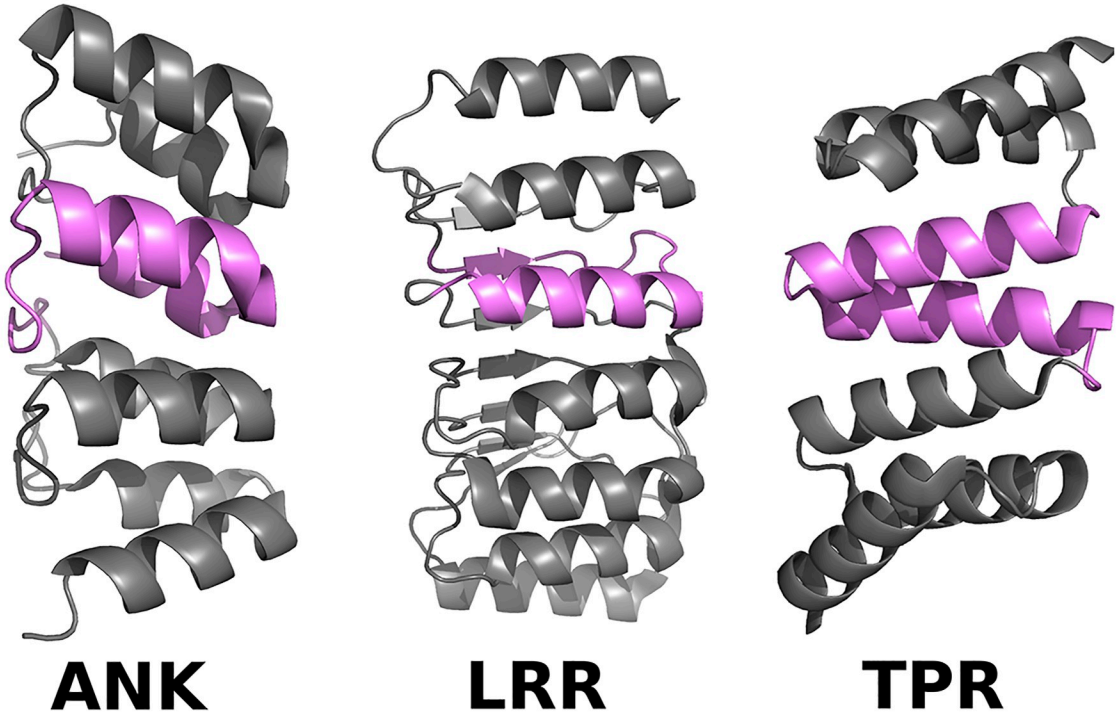

Fig 1. Repeat proteins fold into characteristic accordeon-like folds. Example structures of three protein families are shown, ankyrin repeats (ANK), tetratricopeptide repeats (TPR), leucine-rich repeat (LRR), with the repeating unit highlighted in magenta. All show regular folding patterns with defined contacts in and between repeats.

https://doi.org/10.1371/journal.pcbi.1007282.g001

A recent study [26] addressed the question of the total number of sequences within a given protein family, focusing on ten single-domain families. They took a similar thermodynamic approach to the one followed here, but had to estimate experimentally the free energy threshold ΔG below which the sequences would fold properly. Here we overcome this limitation by forgoing this threshold entirely. Instead we determine the sequence entropy directly, which is argued to be equivalent to using a threshold free energy by virtue of the equivalence of ensembles. We precisely quantify the sequence entropy of three repeat-protein families for which detailed evolutionary energetic fields are known [27]. We explore the properties of the evolutionary landscape shaped by the amino-acid frequency constraints and correlations. We ask whether the energy landscape, defined in sequence space of repeat proteins, is made of a single basin, or rather of a multitude of basins connected by ridges and passes, called "metastable states", as would be expected from spin-glass theory. Using the specific example of repeat proteins makes it possible to analyze the source of the potential landscape ruggedness, and use it to identify which repeat-protein families can be well separated into subfamilies. The rich metastable state structure that we find demonstrates the importance of interactions in shaping the protein family ensemble.

Results

Statistical models of repeat-protein families

We start by building statistical models for the three repeat protein families presented in Fig 1 (ANK, TPR, LRR). These models give the probability $P(\boldsymbol{\sigma})$ to find in the family of interest a particular sequence $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{2L})$ for two consecutive repeats of size L. The model is

designed to be as random as possible, while agreeing with key statistics of variation and co-variation in a multiple sequence alignment of the protein family. Specifically, $P(\boldsymbol{\sigma})$ is obtained as the distribution of maximum entropy [28] which has the same amino-acid frequencies at each position as in the alignment, as well as the same joint frequencies of amino acid usage in each pair of positions. Additionally, repeat proteins share many amino acids between consecutive repeats, both due to sharing a common ancestor and to evolutionary selection acting on the protein. To account for this special property of repeat proteins, we require that the model reproduces the distribution of overlaps $\mathrm{ID}(\boldsymbol{\sigma}) = \sum_{i=1}^{L} \delta_{\sigma_i,\sigma_{i+L}}$ between consecutive repeats. Using the technique of Lagrange multipliers, the distribution can be shown to take the form [17]:

$$P(\boldsymbol{\sigma}) = (1/Z)e^{-E(\boldsymbol{\sigma})}, \tag{1}$$

with

$$E(\boldsymbol{\sigma}) = -\sum_{i=1}^{2L} h_i(\sigma_i) - \sum_{i,j=1}^{2L} J_{ij}(\sigma_i, \sigma_j) + \lambda_{\mathrm{ID}}(\boldsymbol{\sigma}), \tag{2}$$

where $h_i(\sigma)$, $J_{ij}(\sigma_i, \sigma_j)$, and $\{\lambda_{\mathrm{ID}}\}$, ID = 0, 1, . . ., L, are adjustable Lagrange multipliers that are fit to the data to reproduce the experimentally observed site-dependent amino-acid frequencies $f_i(\sigma_i)$, joint probabilities between two positions, $f_{ij}(\sigma_i, \sigma_j)$, and the distribution of Hamming distances between consecutive repeats $P(\mathrm{ID}(\boldsymbol{\sigma}))$, which is equivalent to maximize the likelihood of the data under the model. We fit these parameters using a gradient ascent algorithm: we start from an initial guess of the parameters, then generate sequences via Monte-Carlo simulations and update the parameters proportionally to the difference between the empirical and model generated observables $f_i(\sigma_i) - f_i^{\mathrm{model}}(\sigma_i), f_{ij}(\sigma_i, \sigma_j) - f_{ij}^{\mathrm{model}}(\sigma_i, \sigma_j)$ and $P(\mathrm{ID}(\boldsymbol{\sigma})) - P(\mathrm{ID}(\boldsymbol{\sigma}))^{\mathrm{model}}$. We repeat the previous steps until the model reproduces the empirical observables defined above, with a target precision motivated according to the finite size of our original dataset, as in Ref. [17]. See Methods for more details. We tested the convergence of the model learning by synthetically generating datasets and relearning the model (see Methods).

By analogy with Boltmzan's law, we call $E(\boldsymbol{\sigma})$ a statistical energy, which is in general distinct from any physical energy. The particular form of the energy (2) resembles that of a disordered Potts model. This mathematical equivalence allows for the possibility to study effects that are characteristic of disordered systems, such as frustration or the existence of an energy landscape with multiple valleys, as we will discuss in the next sections.

Eq 2 is the most constrained form of the model, which we will denote by $E_{\mathrm{full}}(\boldsymbol{\sigma})$. One can explore the impact of each constraint on the energy landscape by removing them from the model. For instance, to study the role of inter-repeat sequence similarity due to a common evolutionary origin, one can fit the model without the constraint on repeat overlap ID, *i.e.* without the λ_{ID} term in Eq 2. We call the corresponding energy function E_2. One can further remove constraints on pairwise positions that are not part of the same repeat, making the two consecutive repeats statistically independent and imposing $h_i = h_{i+L}$ (E_{ir}), or only linked through phylogenic conservation through λ_{ID} ($E_{\mathrm{ir},\lambda}$). Finally one can remove all interaction constraints to make all positions independent of each other (E_1), or even remove all constraints ($E_{\mathrm{rand}} \equiv 0$).

Statistical energy *vs* unfolding energy

The evolutionary information contained in multiple sequence alignments of protein families is summarized in our model by the energy function $E(\boldsymbol{\sigma})$. Since this information is often much

easier to access than structural or functional information, there is great interest in extracting functional or structural properties from multiple sequence alignments, provided that there exists a clear quantitative relationship between statistical energy and physical energy.

Such a relationship was determined experimentally for repeat proteins by using $E(\boldsymbol{\sigma})$ to predict the effect of point mutations on the folding stability measured by the free energy difference between the folded and unfolded states, ΔG, called the unfolding energy [17, 20]. Synthetic sequences with low $E(\boldsymbol{\sigma})$ have also been shown to reproduce the fold and function of natural sequences [29]. Here, extending an argument already developed in previous work [30–33], we show how this correspondance between statistical likelihood and folding stability arises in a simple model of evolution.

Evolutionary theory predicts that the prevalence of a particular genotype $\boldsymbol{\sigma}$, i.e. the probability of finding it in a population, is related to its fitness $F(\boldsymbol{\sigma})$. In the limit where mutations affecting the protein are rare compared to the time it takes for mutations to spread through the population, Kimura [34] showed that the probability of a mutation giving a fitness advantage (or disadvantage depending on the sign) ΔF over its ancestor will fix in the population with probability $2\Delta F/(1 - e^{-2N\Delta F})$, where N is the effective population size. The dynamics of successful substitution satisfies detailed balance [35], with the steady state probability

$$P(\boldsymbol{\sigma}) = (1/Z)e^{2NF(\boldsymbol{\sigma})}. \tag{3}$$

Again, one may recognize a formal analogy with Boltzmann's distribution, where F plays the role of a negative energy, and N an inverse temperature. If we now assume that fitness is determined by the unfolding free energy ΔG, $F(\boldsymbol{\sigma}) = f(\Delta G(\boldsymbol{\sigma}))$, then the distribution of genotypes we expect to observe in a population is

$$P(\boldsymbol{\sigma}) = (1/Z)e^{2Nf(\Delta G(\boldsymbol{\sigma}))}. \tag{4}$$

Note that a similar relation should hold even if we relax the hypotheses of the evolutionary model. While in more general contexts (e.g. high mutation rate, recombination), the relation between $\ln P(\boldsymbol{\sigma})$ and $F(\boldsymbol{\sigma})$ may not be linear, such nonlinearities could be subsumed into the function f.

Identifying terms in the two expressions (1) and (3), we obtain a relation between the statistical energy E, and the unfolding free energy ΔG:

$$E(\boldsymbol{\sigma}) = -2Nf(\Delta G(\boldsymbol{\sigma})). \tag{5}$$

For instance, if we assume a linear relation between fitness and ΔG, $f(\Delta G) = A + B\Delta G$, then we get a linear relationship between the statistical energy and ΔG, as was found empirically for repeat proteins [17].

Strikingly, the relationship f does not have to be linear or even smooth for this correspondance to work. Imagine a more stringent selection model, where $f(\Delta G)$ is a threshold function, $f(\Delta G) = 0$ for $\Delta G > \Delta G_{\mathrm{sel}}$ and $-\infty$ otherwise (lethal). In that case the probability distribution is $P(\boldsymbol{\sigma}) = (1/Z)\Theta(\Delta G - \Delta G_{\mathrm{sel}})$, where $\Theta(x)$ is Heaviside's function. Using a saddle-point approximation, one can show that in the thermodynamic limit (long proteins, or large L) the distribution concentrates at the border ΔG_{sel}, and is equivalent to a "canonical" description [30, 31, 33]:

$$P_{\mathrm{sel}}(\boldsymbol{\sigma}) = (1/Z)e^{\Delta G(\boldsymbol{\sigma})/T_{\mathrm{sel}}}, \tag{6}$$

where the "temperature" T_{sel} is set to match the mean ΔG between the two descriptions:

$$\langle \Delta G \rangle_{T_{sel}} = \Delta G_{sel}. \tag{7}$$

This correspondance is mathematically similar to the equivalence between the microcanonical and canonical ensembles in statistical mechanics.

Statistical energy and unfolding free energy are linearly related by equating (Eq 1) and (Eq 6):

$$E(\boldsymbol{\sigma}) = E_0 - \Delta G(\boldsymbol{\sigma})/T_{sel}, \tag{8}$$

despite f being nonlinear. Eq 8 is in fact very general and should hold for any f in the thermodynamic limit in the vicinity of $\langle E \rangle$.

Equivalence between two definitions of entropies

There are several ways to define the diversity of a protein family. The most intuitive one, followed by [26], is to count the total number of amino acid sequences that have an unfolding free energy ΔG_{sel} above a threshold ΔG_{sel} [2]. This number naturally defines a Boltzmann entropy,

$$S = \ln \mathcal{N}(\boldsymbol{\sigma} : \Delta G(\boldsymbol{\sigma}) > \Delta G_{sel}). \tag{9}$$

Alternatively, starting from a statistical model $P(\boldsymbol{\sigma})$, one can calculate its Shannon entropy, defined as

$$S = -\sum_{\sigma} P(\boldsymbol{\sigma}) \ln P(\boldsymbol{\sigma}), \tag{10}$$

as was done in Ref. [27]. What is the relation between these two definitions?

By the same saddle-point approximation as in the previous section, the two are identical in the thermodyamic limit (large L), provided that the condition (Eq 7) is satisfied. We can thus reconcile the two definitions of the entropy in that limit.

To calculate the Boltzmann entropy (Eq 9), one needs to first evaluate the threshold E_{sel} in terms of statistical energy. This threshold is given by $E_{sel} = E_0 - \Delta G_{sel}/T_{sel}$, where E_0 and T_{sel} can be obtained directly by fitting (Eq 8) to single-mutant experiments. E_{sel} can also be obtained as a discrimination threshold separating sequences that are known to fold properly versus sequences that do not [26]. In that case, assuming that the linear relationship (Eq 8) was evaluated empirically using single mutants, this relationship can be inverted to get ΔG_{sel} in physical units.

Calculating the Shannon entropy Eq (10), on the other hand, does not require to define any threshold. However, the threshold in the equivalent Boltzmann entropy can be obtained using Eqs 7 and 8, i.e. $E_{sel} = \langle E \rangle$, where the average is performed using the distribution defined in Eqs 1 and 2.

Entropy of repeat protein families

To compare how the different elements of the energy function affect diversity, we calculate the entropy of ensembles built of two consecutive repeats from a given protein family for the different kinds of models described earlier, from the least constrained to the most constrained: E_{rand}, E_1, E_{ir}, $E_{ir,\lambda}$, E_2, E_{full}. In the case of models with interactions, calculating the entropy directly from the definition Eq (10) is impossible due to the large sums. A previous study of entropies of protein families used an approximate mean-field algorithm, called the Adaptive

Table 1. Entropies (in bits, i.e. units of ln(2)) of sequences made of two consecutive repeats, for the three protein families shown in Fig 1. Entropies are calculated for models of different complexity: model of random amino acids ($S_{rand} = 2L \ln(21)$, divided by ln(2) when expressed in bits); independent-site model (S_1), pairwise interaction model (S_2); pairwise interaction model with constraints due to repeat similarity λ_{ID} (S_{full}); pairwise interaction model of two non-interacting repeats learned without (S_{ir}) and with ($S_{ir,\lambda}$) constraints on repeat similarity. Fig 2 shows graphically some of the information contained in this table.

family	$2L$	S_{rand}	S_1	S_2	S_{full}	S_{ir}	$S_{ir,\lambda}$
ANK	66	290	181 ± 0.05	169.7 ± 0.6	167.2 ± 0.3	176.7 ± 0.1	172 ± 0.4
LRR	48	211	130 ± 0.05	114 ± 0.4	113.2 ± 0.3	123.1 ± 0.1	118.8 ± 0.1
TPR	68	299	169 ± 0.1	145.4 ± 0.7	141.4 ± 0.3	157.6 ± 0.1	146.9 ± 0.4

https://doi.org/10.1371/journal.pcbi.1007282.t001

Cluster Expansion [27], for both parameter fitting and entropy estimation. Here we estimated the entropies using thermodynamic integration of Monte-Carlo simulations, as detailed in Methods. This method is expected to be asymptotically unbiased and accurate in the limit of large Monte-Carlo samples.

The resulting entropies and their differences are reported in Table 1 and Fig 2. All three considered families (ankyrins (ANK), leucine-rich repeats (LRR), and tetratricopeptides (TPR)) show a large reduction in entropy ($\sim 40 - 50\%$) compared to random polypeptide string models of the same length $2L$ (of entropy $S_{rand} = 2L \ln(21)$). Interactions and phylogenic similarity between repeats generally have a noticeable effect on family diversity, although the magnitude of this effect depends on the family: $(S_1 - S_{full})/S_{full} = 7\%$ for ANK, versus, 13% for LRR, and 16% for TPR. Thus, although interactions are essential in correctly predicting the folding properties, they seem to only have a modest effect on constraining the space of accessible proteins compared to that of single amino-acid frequencies. However, when converted to numbers of sequences, this reduction is substantial, from $e^{S_1} \sim 3 \cdot 10^{54}$ to $e^{S_{full}} \sim 2 \cdot 10^{50}$ for ANK, from 10^{39} to 10^{34} for LRR, and from $7 \cdot 10^{50}$ to $4 \cdot 10^{42}$ for TPR.

By considering models with more and more constraints, and thus with lower and lower entropy, we can examine more finely the contribution of each type of correlation to the entropy reduction, going from E_1 to E_{ir} to $E_{ir,\lambda}$ to E_{full}. This division allows us to quantify the relative importance of phylogenic similarity between consecutive repeats (λ_{ID}) relative to the impact of functional interactions (J_{ij}), as well as the relative weights of repeat-repeat versus within-repeat interactions (Fig 2). We find that phylogenic similarity contributes substantially to the entropy reduction, as measured by $S_{ir} - S_{ir,\lambda} = 4.5$ bits for ANK, 4.3 bits for LRR, and 10.7 bits for TPR. The contribution of repeat-repeat interactions ($S_{ir,\lambda} - S_{full} \sim 5$ bits for all three families) is comparable or of the same order of magnitude as that of within-repeat interactions ($S_1 - S_{ir} = 4.3$ bits for ANK, 6.9 bits for LRR, and 11.4 bits for TPR). This result emphasizes the importance of physical interactions between neighboring repeats in the whole protein.

On a technical note, we also find that pairwise interactions encode constraints that are largely redundant with the constraint of phylogenic similarity between consecutive repeats, as can be measured by the double difference $S_{ir} - S_{ir,\lambda} - S_2 + S_{full} > 0$ (Fig 2, orange bars). This redundancy comes from the fact that, in absence of an explicit constaint on $P(\mathrm{ID})$ in E_2, the interaction couplings $J_{i,i+L}(\sigma, \sigma)$ between homologous positions in the two repeats is expected to favor pairs of identical residues to mimic the effect of λ_{ID}. This redundancy motivates the need to correct for this phylogenic bias before estimating repeat-repeat interactions.

Comparing the three families, ANK has little phylogenic bias between consecutive repeats, and relatively weak interactions. By contrast, TPR has a strong phylogenic bias and strong within-repeat interactions.

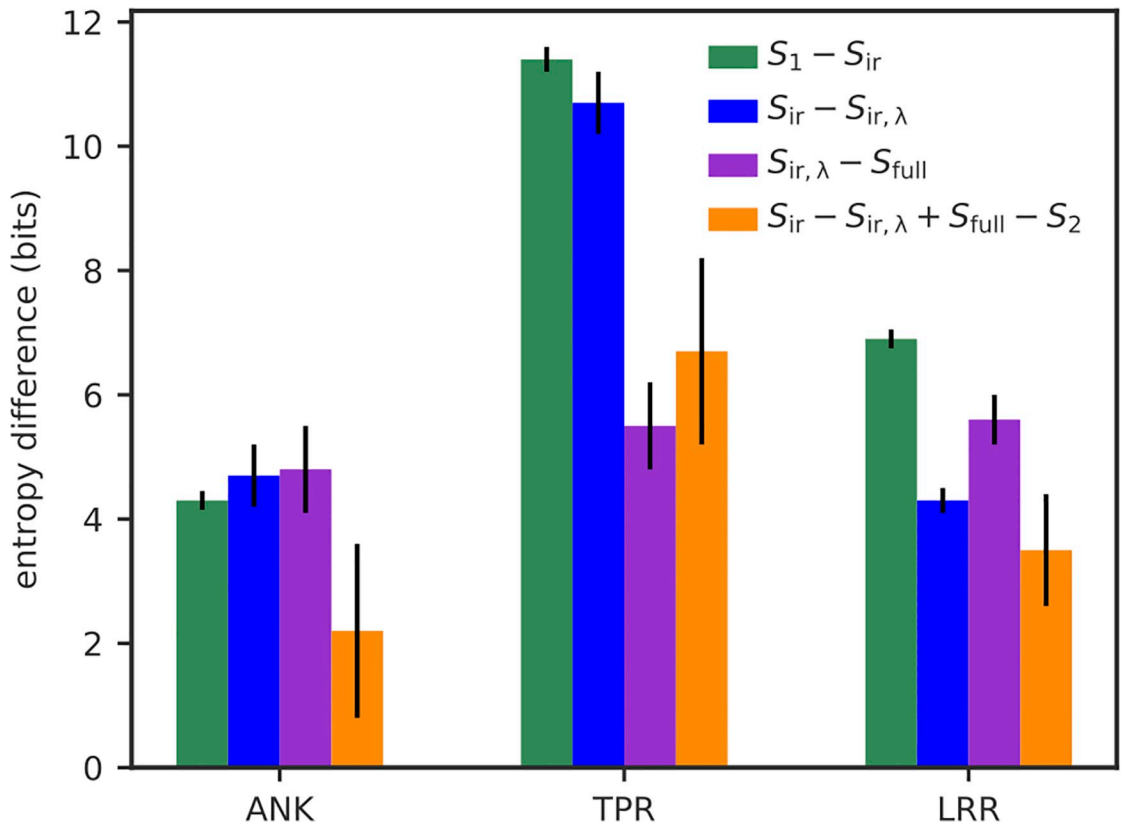

Fig 2. Contributions of within-repeat interactions ($S_1 - S_{ir}$ green), repeat-repeat interactions ($S_{ir,\lambda} - S_{full}$, purple), and phylogenic bias between consecutive repeats ($S_{ir} - S_{ir,\lambda}$, blue), to the entropy reduction from an independent-site model. All three contributions are comparable, but with a larger effect of within-repeat interactions and phylogenic bias in TPR. The fourth bar (orange) quantifies the redundancy between two constraints with overlapping scopes: the constraint on consecutive-repeat similary, and the constraint on repeat-repeat correlations. This redundancy is naturally measured within information theory by the difference of impact (i.e. entropy reduction) of a constraint depending on whether or not the other constraint is already enforced.

https://doi.org/10.1371/journal.pcbi.1007282.g002

Effect of interaction range

We wondered whether interactions constraining the space of accessible proteins had a characteristic lengthscale. To answer this question, for each protein family in Fig 1, we learn a sequence of models of the form Eq 2, in which J_{ij} was allowed to be non-zero only within a certain interaction range $d(i, j) \leq W$, where the distance $d(i, j)$ between sites i and j can be defined in two different ways: either the linear distance $|i - j|$ expressed in number of amino-acid sites, or the three-dimensional distance between the closest heavy atoms in the reference structure of the residues. Details about the learning procedure and error estimation are given in the Methods; see also S1 Fig for an alternative error estimate.

The entropy of all families decreases with interaction range W, both in linear and three-dimensional distance, as more constraints are added to reduce diversity (Fig 3 for ANK, and S2 Fig for LRR and TPR). The initial drop as a function of linear distance (Fig 3A) is explained by the many local interactions between nearby residues in the sequence. The entropy then plateaus until interactions between same-position residues in consecutive repeats are included in the W range, which leads to a sharp entropy drop at $W = L$. This suggests that long range interactions along the sequence generally do not constrain the protein ensemble diversity, except for interactions at exactly the scale of the repeat. This result suggests that the repeat structure is

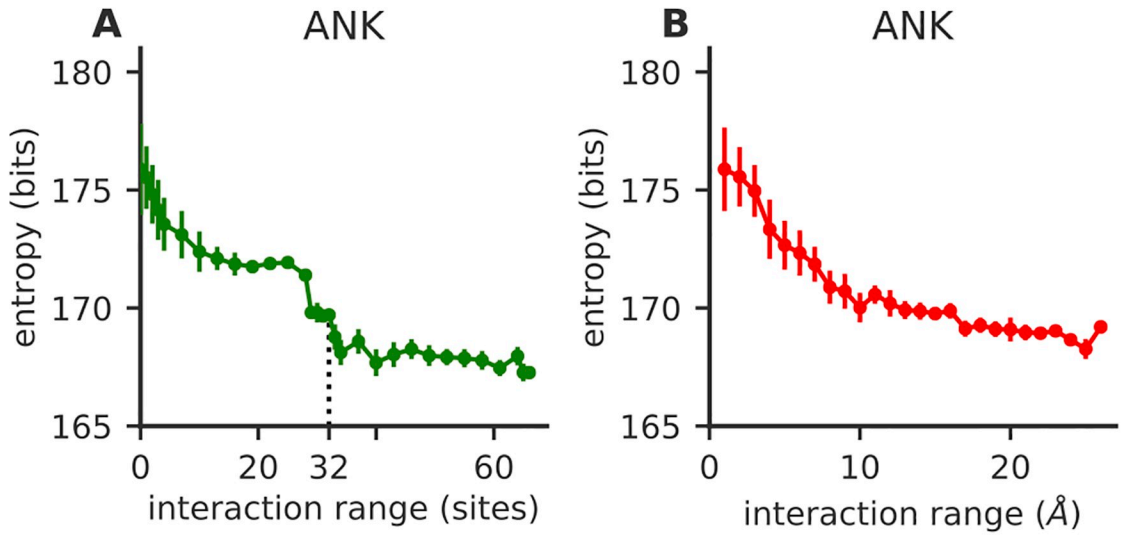

Fig 3. Entropy reduction as a function of the range of interactions between residue sites. A) Entropy of two consecutive ANK repeats, as a function of the maximum allowed interaction distance W along the linear sequence. The entropy of the model decreases as more interactions are added and they constrain the space of possible sequences. After a sharp initial decrease at short ranges, the entropy plateaus until interactions between complementary sites in neighbouring repeats lead to a secondary sharp decrease at $W = L - 1 = 32$ (dashed line), due to structural interactions between consecutive repeats. B) Entropy of two consecutive ANK repeats as a function of the maximum allowed three-dimensional interaction range. The entropy decreases rapidly until ~ 10 Angstrom, after which decay becomes slower. In both panels entropies are averaged over 10 realizations of fitting the model; See Methods for details of the learning and entropy estimation procedure. Error bars are estimated from fitting errors between the data and the model; see Methods and S1 Fig for error bars calculated as standard deviations over 10 realizations of model fitting.

https://doi.org/10.1371/journal.pcbi.1007282.g003

an important constraint limiting protein sequence exploration. These observations hold for all three repeat protein families. The importance of 3D structure in reducing the entropy can also be appreciated in the entropy decay as a function of physical distance (Fig 3B for ANK) where most of the entropy drop happens within the first 10 angstroms, indicating that above this characteristic distance interactions are not crucial in constraining the space of accessible sequences.

Multi-basin structure of the energy landscape

The energy function of Eq (2) takes the same mathematical form as a disordered Potts model. These models, in particular in cases where σ_i can only take two values, have been extensively studied in the context of spin glasses [36]. In these systems, the interaction terms $-J_{ij}(\sigma_i, \sigma_j)$ imply contradictory energy requirements, meaning that not all of these terms can be minimized at the same time—a phenomenon called frustration. Because of frustration, natural dynamics aimed at minimizing the energy are expected to get stuck into local, non-global energy minima (Fig 4), significantly slowing down thermalization. This phenomenon is similar to what happens in structural glasses in physics, where the energy landscape is "rugged" with many local minima that hinder the dynamics. Incidentally, concepts from glasses and spin glasses have been very important for understanding protein folding dynamics [37].

We asked whether the energy landscape of Eq (2) was rugged with multiple minima, and investigated its structure. To find local minima, we performed a local energy minimization of E_{full} (learned with all constraints including on $P(\text{ID})$, but taken with $\lambda_{\text{ID}} = 0$ to focus on functional energy terms). By analogy with glasses, such a minimization is sometimes called a zero-temperature Monte-Carlo simulation or a "quench". The minimization procedure was started

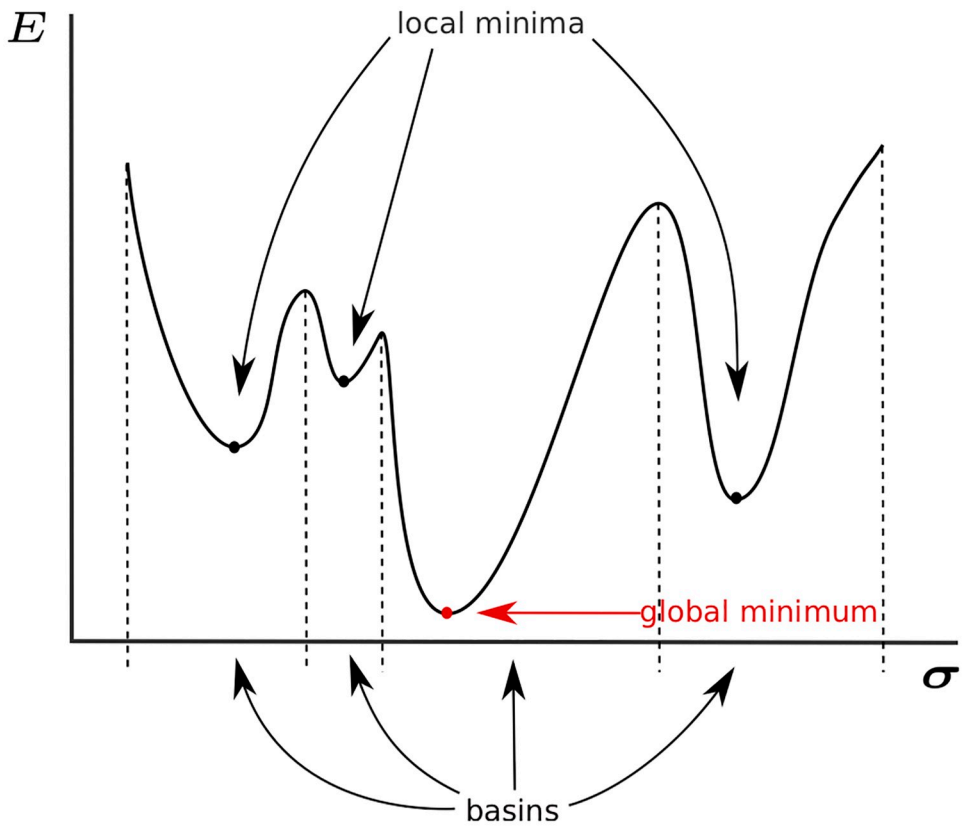

Fig 4. A rugged energy landscape is characterized by the presence of local minima, where proteins sequences can get stuck during the evolutionary process. The set of sequences that evolve to a given local minimum defines the *basin of attraction* of that mimimum.

https://doi.org/10.1371/journal.pcbi.1007282.g004

from many initial conditions corresponding to naturally occuring sequences of consecutive repeat pairs. At each step of the minimization, a random beneficial (energy decreasing) single mutation is picked; double mutations are allowed if they correspond to twice the same single mutation on each of the two repeats. Minimization stops when there are no more beneficial mutations. This stopping condition defines a local energy minimum, for which any mutation increases the energy. The set of sequences which, when chosen as initial conditions, lead to a given local minimum defines the *basin of attraction* of that energy mimimum (Fig 4). The size of a basin corresponds to the number of natural proteins belonging to that basin.

Performing this procedure on natural sequences of consecutive repeat pairs from all three families yielded a large number of local minima (Fig 5). To control for the phylogenetic bias that links natural sequences, we repeated this analysis on sequences synthetically generated from the model (E_{full}), and obtained very similar results (see S6 Fig for ANK). When ranked from largest to smallest, the distribution of basin sizes follows a power law (Fig 5A for ANK and S3A and S4A Figs for LRR and TPR). The energy of the minimum of each basin generally increases with the rank, meaning that largest basins are also often the lowest. Despite this multiplicity of local minima, the Monte-Carlo dynamics that we used in previous sections for learning the model parameters and for estimating the entropy did not get stuck in these minima, suggesting relatively low energy barriers between them.

The partition of sequences into basins allows for the definition of a new kind of entropy $S_{\text{conf}} = -\Sigma_b P(b) \ln P(b)$ called configurational entropy, based on the distribution of basin sizes,

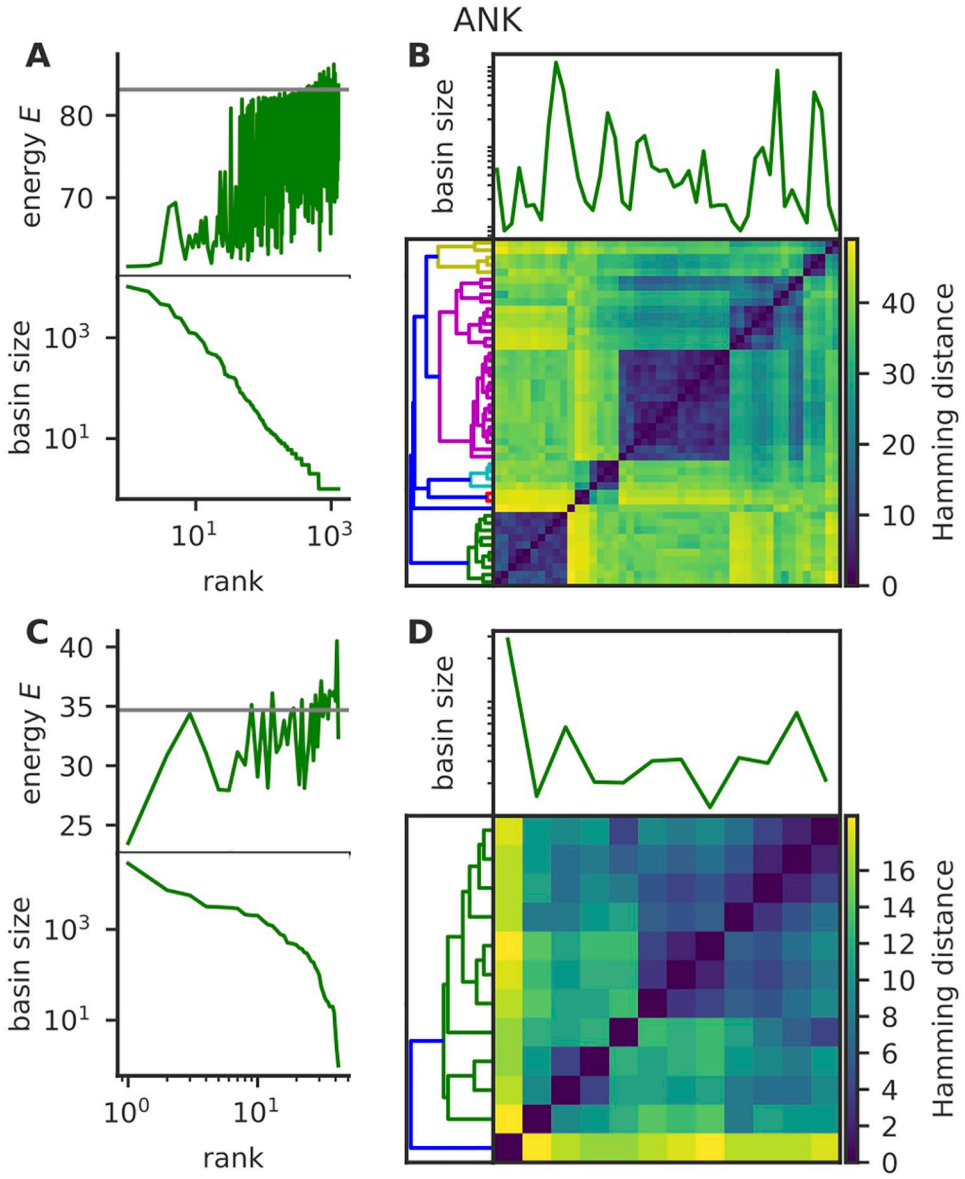

Fig 5. Interactions within and between repeats sculpt a rugged energy landscape with many local minima. Local minima were obtained by performing a zero-temperature Monte-Carlo simulation with the energy function in Eq (2), starting from initial conditions corresponding to naturally occurring sequences of pairs of consecutive ANK repeats. A, bottom) Rank-frequency plot of basin sizes, where basins are defined by the set of sequences falling into a particular minimum. A, top) energy of local minima *vs* the size-rank of their basin, showing that larger basins often also have lowest energy. Gray line indicates the energy of the consensus sequence, for comparison. B) Pairwise distance between the minima with the largest basins (comprising 90% of natural sequences), organised by hierarchical clustering. The panel right above the matrix shows the size of the basins relative to the minima corresponding to the entries of the distance matrix. A clear block structure emerges, separating different groups of basins with distinct sequences. C-D) Same as A) and B) but for single repeats. Since single repeats are shorter than pairs (length L instead of $2L$), they have fewer local energy minima, yet still show a rich multi-basin structure. Equivalent analyses for LRR and TPR are shown in S3 and S4 Figs.

https://doi.org/10.1371/journal.pcbi.1007282.g005

$P(b) = \sum_{\boldsymbol{\sigma}\in b} P(\boldsymbol{\sigma})$, where $\boldsymbol{\sigma} \in b$ means that energy minimization starting with sequence $\boldsymbol{\sigma}$ leads to basin b. This configurational entropy measures the effective diversity of basins, and is thus much lower than the sequence entropy S_{full}, while the difference $S_{\text{full}} - S_{\text{conf}}$ measures the average diversity of sequences within each basin. We find $S_{\text{conf}} = 5.1$ bits for ANK, 6.0 bits for LRR,

and 10.4 bits for TPR. As each basin corresponds to a distinct sub-family within each family [32], this entropy quantifies the effective number of these subgroups.

While basins are very numerous, they are also not independent of each other. An analysis of pairwise distances (measured as the Hamming distance between the local minima) between the largest basins reveals that they can be organised into clusters (panels B of Fig 5, S3 and S4 Figs), suggesting a hierarchical structure of basins, as is common in spin glasses [36].

The impact of repeat-repeat interactions on the multi-basin structure can be assessed by repeating the analysis on the model of non-interacting repeats, E_{ir}. In that model the two repeats are independent, so it suffices to study local energy minima of single repeats—local minima of pairs of repeats follow simply from the combinatorial pairing of local minima in each repeat. The analyses of basin size distributions, energy minima, and pairwise distances in single repeats are shown in panels C and D of Fig 5, S3, and S4 Figs. We still find a substantial number of unrelated energy minima, suggesting again several distinct subfamilies even at the single-repeat level. For comparison, the configurational entropy of pairs of independent repeats is 6.9 bits for ANK, 6.7 for LRR, and 7.6 for TPR. While for ANK and LRR repeat-repeat interactions decrease the configurational entropy, as they do for the conventional entropy, they in fact increase entropy for TPR, making the energy landscape even more frustrated and rugged.

Note that the independent sites model E_1 defines a convex energy landscape with a single local minimum—the consensus sequence—as all constraints h_i can be optimized independently. To address how the interactions contribute in shaping the sequence space, going from a convex to a rugged landscape, we repeated the analysis with a limited linear interaction range W of 3 and 10 (models of Fig 3A). We find that the more interactions we add, the more local minima we find (S5A and S5B Fig for ANK with $W = 3$, and C and D for $W = 10$). The minima cluster into clearer sub-blocks structure as the interaction range is increased, consistent with the entropy reduction observed in Fig 3A.

In summary, the analysis of the energy landscape reveals a rich structure, with many local minima ranging many different scales, and with a hierarchical structure between them.

Distance between repeat families

Lastly, we compared the statistical energy landscapes of different repeat families. Specifically, we calculated the Kullback-Leibler divergence between the probability distributions $P(\boldsymbol{\sigma})$ (given by Eqs 1 and 2) of two different families, after aligning them together in a single multiple sequence alignment (see Methods).

We find essentially no similarity between ANK and TPR, despite them having similar lengths: $D_{\mathrm{KL}}(\mathrm{ANK}||\mathrm{TPR}) = 227.6$ bits, and $D_{\mathrm{KL}}(\mathrm{TPR}||\mathrm{ANK}) = 214.1$ bits. These values are larger than the Kullback-Leibler divergence between the full models for these families and a random polypetide, $D_{\mathrm{KL}}(\mathrm{ANK}||\mathrm{rand}) = 122.8$ bits, and $D_{\mathrm{KL}}(\mathrm{TPR}||\mathrm{rand}) = 157.6$ bits. LRR is not comparable to ANK or TPR as it is much shorter, and a common alignment is impractical. These large divergences between families of repeat proteins show that different families impose quantifiably different constraints, which have forced them to diverge into different troughs of non-overlapping energy landscapes. This lack of overlap makes it impossible to find intermediates between the two families that could evolve into proteins belonging to both families.

Discussion

Our analysis of repeat protein families shows that the constraints between amino acids in the sequences allows for an estimation of the size of the accessible sequence space. The obtained numbers (ranging from 141 bits to 167 bits, corresponding to 10^{36} to 10^{50} sequences) are of

course huge compared to the number of sequences in our initial samples ($\sim 20,500$ for ANK, $\sim 18,800$ for LRR, and $\sim 10,000$ for TPR), but comparable to the total number of proteins having been explored over the whole span of evolution, estimated to be 10^{43} in Ref. [1].

In particular, we have quantified the reduction of the accessible sequence space with respect to random polypeptides. While most of this reduction is attributable to conservation of residues at each site, interactions between amino acids, both within and between consecutive repeats, significantly constrain the diversity of all repeat families. The break-up of entropy reduction between the three different sources of constraints—within-repeat interactions, between-repeat interactions, and evolutionary conservation between consecutive repeats—is fairly balanced, although TPR stands out as having more within-repeat interactions and more conservation between neighbours, suggesting that it may have had less time to equilibrate.

All studied repeat families have rugged energy landscapes with multiple local energy minima. Note that the emergence of this multi-valley landscape is a consequence of the interactions between amino acids: models of independent positions (E_1) only admit a single energy minimum corresponding to the consensus sequence. This multiplicity of minima allow us to collapse multiple sequences to a small number of coarse-grained attractor basins. These basins suggest that mutations between sequences within one coarse-grained basin are much more likely than mutating into sequences in other basins. In general, our results paint a picture of further subdivisions within a family, and define sub-families due to the fine grained interaction structure. Going beyond single families, this analysis suggest a view in which natural proteins all live in a global evolutionary landscape, of which families would be basins, or clusters of basins, with a hierarchical structure [32].

This overall picture of the sequence energy landscape is reminiscent of the hierarchical picture of the structural energy landscape of globular proteins, an overall funneled shape with tiers within tiers [38]. The form of the energy landscape forcibly shapes the accessible evolutionary paths between sequences. The rugged and further subdivided structure shows that the uncovered constraints are global, and not just pairwise between specific residues. Therefore even changing two residues together, as is often done in laboratory experiments, is not enough to recover the evolutionary trajectories. While other approaches have explored local accessible directions of evolution [39], our results suggest more global, non local modes of evolution between clusters.

Interestingly, the sequences that correspond to the energy minima of the landscapes are not found in the natural dataset. This observation can be either due to sampling bias (we have not yet observed the sequence with the minimal energy, although it exists), or this sequence may not have been sampled by nature. Alternatively, there may be additional functional constraint that are not included in our model to avoid these low energy sequences (e.g. a too stable protein may be difficult to degrade).

Even more intriguingly, sequences with minimal energy do not correspond to the consensus sequence of the alignment (whose energy is marked by a gray line in panel A of Fig 5, S3, and S4 Figs), suggesting that the consensus sequence can be improved upon. All three repeat protein families studied here have been shown to be amenable to simple consensus-guided design of synthetic proteins. Synthetic proteins based on the consensus sequences of multiple alignments [40] were found to be foldable and very stable against chemical and thermal denaturation. Mutations towards consensus amino acids in the ANK family members have been experimentally shown to both stabilize the whole repeat-array and they may tune the folding paths towards nucleating folding in the consensus sites [41, 42]. Our results suggest that interactions may play an additional role in stabilizing the sequences, and propose alternative solutions to the consensus sequences in the design of synthetic proteins.

Methods

Data curation

We use a previously curated alignment of pairs of repeats for each family [17]: ANK (PFAM id PF00023 with a final alignment of 20513 sequences of $L = 66$ residues each), LRR (PFAM id PF13516 with a final alignment of 18839 sequences of $L = 48$ residues each) and TPR (PFAM id PF00515 with a final alignment of 10020 sequences of $L = 68$ residues each). Those multiple sequence alignments of repeats were obtained from PFAM 27.0 [6, 7]. In order to improve the data obtained from the PFAM database, we used original full protein sequences available in UniProt database [43] to add available information using the headers of the original alignement. Firstly, to decrease the number of gaps positions, misdetected initial and final amino acids in repeats were completed with residues from full sequences. Secondly, individual repeats which appeared consecutively in natural proteins were joined into pairs. Finally, positions with more than 80% of gaps along the alignment were removed, eliminating in this way insertions.

From the multiple sequence alignement of each family, they were calculated the observables that we use to constrain our statistical model. Particularly, we calculated the marginal frequency $f_i(\sigma_i)$ of an amino acid σ_i at position i and the joint frequency $f_{ij}(\sigma_i, \sigma_j)$ of two amino acids σ_i and σ_j at two different positions i and j. These quantities were calculated using only sequences selected by clustering at 90% of identity computed with CD-HIT [44] and then normalizing by the amount of sequences. In this way, the occurrences of residues in every position are not biased by overrepresentation of proteins in the database. Furthermore, to take into account the repeated nature of the protein families that we are considering, an additional observable was calculated, the distribution of sequence overlap between two consecutive repeats, $P(\mathrm{ID}(\boldsymbol{\sigma}))$, with $\mathrm{ID}(\boldsymbol{\sigma}) = \sum_{i=1}^{L} \delta_{\sigma_i,\sigma_{i+L}}$.

Model fitting

In order to obtain a model that reproduces the experimentally observed site-dependent amino-acid frequencies, $f_i(\sigma_i)$, correlations between two positions, $f_{ij}(\sigma_i, \sigma_j)$, and the distribution of Hamming distances between consecutive repeats, $P(\mathrm{ID}(\boldsymbol{\sigma}))$, we apply a likelihood gradient ascent procedure, starting from an initial guess of the $h_i(\sigma_i)$, $J_{ij}(\sigma_i, \sigma_j)$ and $\lambda_{\mathrm{ID}}(\boldsymbol{\sigma})$ parameters.

At each step, we generate 80000 sequences of length $2L$ through a Metropolis-Hastings Monte-Carlo sampling procedure. We start from a random amino-acid sequence and we produce many point mutations in any position, one at a time. If a mutation decreases the energy (2) we accept it. If not, we accept the mutation with probability $e^{-\Delta E}$, where ΔE is the difference of energy between the original and the mutated sequence. We add one sequence to our final ensemble every 1000 steps. Once we generated the sequence ensemble, we measure its marginals $f_i^{\mathrm{model}}(\sigma_i)$ and $f_{ij}^{\mathrm{model}}(\sigma_i, \sigma_j)$, as well as $P^{\mathrm{model}}(\mathrm{ID}(\boldsymbol{\sigma}))$, and update the parameters of Eq 2 following the gradient of the likelihood. The local field and $\lambda_{\mathrm{ID}}(\boldsymbol{\sigma})$ are updated along the gradient of the per-sequence log-likelihood, equal to the difference between model and data averages:

$$h_i(\sigma_i)^{t+1} \leftarrow h_i(\sigma_i)^t + \epsilon_m[f_i(\sigma_i) - f_i^{\mathrm{model}}(\sigma_i)], \tag{11}$$

$$\lambda_{\mathrm{ID}}(\boldsymbol{\sigma})^{t+1} \leftarrow \lambda_{\mathrm{ID}}(\boldsymbol{\sigma})^t - \epsilon_{\mathrm{ID}}[P(\mathrm{ID}(\boldsymbol{\sigma})) - \mathrm{P}(\mathrm{ID}(\boldsymbol{\sigma}))^{\mathrm{model}}]. \tag{12}$$

As the number of parameters for the interaction terms J_{ij} is large ($= 21^2 L^2$), we force to 0 those that are not contributing significantly to the model frequencies through a L_1

regularisation $\gamma\Sigma_{ij,\,\sigma,\,\tau}|J_{ij}(\sigma,\tau)|$ added to the likelihood. This leads to the following rules of maximization: If $J_{ij}(\sigma_i,\sigma_j)^t = 0$ and $|f_{ij}(\sigma_i,\sigma_j) - f_{ij}^{\text{model}}(\sigma_i,\sigma_j)| < \gamma$

$$J_{ij}(\sigma_i,\sigma_j)^{t+1} \leftarrow 0. \tag{13}$$

If $J_{ij}(\sigma_i,\sigma_j)^t = 0$ and $|f_{ij}(\sigma_i,\sigma_j) - f_{ij}^{\text{model}}(\sigma_i,\sigma_j)| > \gamma$

$$J_{ij}(\sigma_i,\sigma_j)^{t+1} \leftarrow \epsilon_j[f_{ij}(\sigma_i,\sigma_j) - f_{ij}^{\text{model}}(\sigma_i,\sigma_j) - \gamma\,\text{sign}(f_{ij}(\sigma_i,\sigma_j) - f_{ij}^{\text{model}}(\sigma_i,\sigma_j))]. \tag{14}$$

If $[J_{ij}(\sigma_i,\sigma_j)^t + \epsilon_j[f_{ij}(\sigma_i,\sigma_j) - f_{ij}^{\text{model}}(\sigma_i,\sigma_j) - \gamma\,\text{sign}(J_{ij}(\sigma_i,\sigma_j)^t)]]J_{ij}(\sigma_i,\sigma_j)^t \geq 0$

$$J_{ij}(\sigma_i,\sigma_j)^{t+1} \leftarrow J_{ij}(\sigma_i,\sigma_j)^t + \epsilon_j[f_{ij}(\sigma_i,\sigma_j) - f_{ij}^{\text{model}}(\sigma_i,\sigma_j) - \gamma\,\text{sign}(J_{ij}(\sigma_i,\sigma_j)^t)]. \tag{15}$$

If $[J_{ij}(\sigma_i,\sigma_j)^t + \epsilon_j[f_{ij}(\sigma_i,\sigma_j) - f_{ij}^{\text{model}}(\sigma_i,\sigma_j) - \gamma\,\text{sign}(J_{ij}(\sigma_i,\sigma_j)^t)]]J_{ij}(\sigma_i,\sigma_j)^t < 0$

$$J_{ij}(\sigma_i,\sigma_j)^{t+1} \leftarrow 0. \tag{16}$$

The optimization parameters were set to: $\epsilon_m = 0.1$, $\epsilon_j = 0.05$, $\epsilon_{\text{ID}} = 10$, and $\gamma = 0.001$.

To estimate the model error, we compute $f_i(\sigma_i) - f_i^{\text{model}}(\sigma_i)$ and $f_{ij}(\sigma_i,\sigma_j) - f_{ij}^{\text{model}}(\sigma_i,\sigma_j)$. We also calculate the difference of generated and natural repeat similarity distribution for all the possible repeats Hamming distances, penalized by a factor 5 to better learn the parameter λ_{ID}: $5(P(\text{ID}(\boldsymbol{\sigma})) - P(\text{ID}(\boldsymbol{\sigma}))^{\text{model}})$. We repeat the procedure above until the maximum of all errors, $|f_i(\sigma_i) - f_i^{\text{model}}(\sigma_i)|$, $|f_{ij}(\sigma_i,\sigma_j) - f_{ij}^{\text{model}}(\sigma_i,\sigma_j)|$ and $5|P(\text{ID}(\boldsymbol{\sigma})) - P(\text{ID}(\boldsymbol{\sigma}))^{\text{model}}|$, goes below 0.02, as in Ref. [17].

Models with different sets of constraints

Using this procedure we can calculate the model defined in Eq 2 with different interaction ranges used in the entropy estimation in Fig 3A. We start from the independent model $h_i(\sigma_i) = \log f_i(\sigma_i)$. We first learn the model in Eq 2 with $J = 0$. We then re-learn models with interactions between sites i, j along the linear sequence such that $|i - j| \leq W$, in a seeded way starting from the previous model. The first and last point of Fig 3 correspond to the independent site model with λ_{ID} and the full model in Eq 2

The entropy in Fig 3B is calculated in the same way as in Fig 3, but now interactions are turned on progressively according to physical distance in the 3D structure rather than the linear sequence distance. In order to obtain the physical distance between residues we use as a reference structure the first two repeats of a consensus designed ankyrin protein 1n0r [45, 46], which have exactly 66 amino-acids. We define the 3D separation between two residues as the minimum distance between their heavy atoms in the reference structure.

To learn the Potts model without λ_{ID} (E_2) we remove λ_{ID} from Eq 2 and re-learn the Potts field using the full model parameters as initial contition.

To learn the single repeat models with and without λ (E_{ir} and $E_{\text{ir},\lambda}$, we take as initial condition the model with interactions below the length of a repeat ($W = L - 1$, dashed vertical line in Fig 3), and then learn a model removing all the J_{ij} terms between different repeats. We also impose that the h_i fields and intra-repeats J_{ij} terms are the same in each repeat, and the experimental amino-acid frequencies to be reproduced by the model are the average over the two repeats of the 1- and 2-points intra-repeats frequencies $f_i(\sigma_i)$ and $f_{ij}(\sigma_i,\sigma_j)$, such that

$$f_i'(\sigma_i) = f_{i+L}'(\sigma_i) = \frac{1}{2}\left(f_i(\sigma_i) + f_{i+L}(\sigma_i)\right), \tag{17}$$

and

$$\begin{aligned} f'_{ij}(\sigma_i, \sigma_j) &= f'_{i+L,j+L}(\sigma_i, \sigma_j) = \\ &= \frac{1}{2}(f_{ij}(\sigma_i, \sigma_j) + f_{i+L,j+L}(\sigma_i, \sigma_j)), \end{aligned} \tag{18}$$

if i and j represent sites within the same repeat. In this way we obtain a model for a single repeat that can be extended to both the repeats in the original set of sequences of our dataset.

Entropy estimation

In practice to calculate the entropy S of the protein families we relate it to the internal energy $E = -\log p(\boldsymbol{\sigma})$ and the free energy $F = -\log Z$:

$$\begin{aligned} S &= \langle E \rangle - F \\ &= \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) E(p(\boldsymbol{\sigma})) + \log Z \\ &= -\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \log p(\boldsymbol{\sigma}) , \end{aligned} \tag{19}$$

We generate sequences according to the energy function in Eq 2 and use them to numerically compute $\langle E \rangle$. To calculate the free energy we use the auxilliary energy function:

$$E_\alpha(\boldsymbol{\sigma}) = -\sum_i h_i(\sigma_i) + \alpha \left[-\sum_{ij} J_{ij}(\sigma_i, \sigma_j) + \lambda_{\mathrm{ID}} \right] , \tag{20}$$

where the interaction strength across different sites can be tuned through a parameter α that is changed from 0 to 1. We generate protein sequence ensembles with different values of α and use them to calculate F as a function of α, $F(1) = F(0) + \int_0^1 \mathrm{d}\alpha \frac{\mathrm{d}F}{\mathrm{d}\alpha}$:

$$F(1) = F(0) + \int_0^1 \mathrm{d}\alpha \left\langle -\sum_{ij} J_{ij}(\sigma_i, \sigma_j) + \lambda_{\mathrm{ID}} \right\rangle_\alpha , \tag{21}$$

where the average over α is taken over the sequences generated with a certain value of α, characterized by the ensemble with probability $p_\alpha(\boldsymbol{\sigma}) = (1/Z_\alpha) e^{-E_\alpha(\boldsymbol{\sigma})}$. $F(0)$ is the free energy for an independent sites model:

$$F(0) = -\sum_i \log \sum_{\sigma_i} \mathrm{e}^{h_i(\sigma_i)} , \tag{22}$$

where the first sum is taken over protein sites and the second over all possible amino-acids at a given site. Eqs 22 and 19 result in the thermodynamic sampling approximation for calculating the entropy [47]:

$$S = \langle E \rangle + \sum_i \log \sum_{\sigma_i} \mathrm{e}^{h_i(\sigma_i)} - \int_0^1 \mathrm{d}\alpha \left\langle -\sum_{ij} J_{ij}(\sigma_i, \sigma_j) + \lambda_{\mathrm{ID}} \right\rangle_\alpha . \tag{23}$$

We generate 80000 sequences using Monte Carlo sampling for the energy in Eq 20 with 50 different α values, equally spaced between 0 and 1 at a distance of 0.02, and then numerically compute the integral in Eq 23 using the Simpson rule.

Entropy error

The entropy estimate is subject to three sources of uncertainty: the finite-size of the dataset, convergence of parameter learning, and the noise in the thermodynamic integration. We estimate the contribution of each of these errors using the independent sites model. In the independent sites model each site i is simply described by a multinomial distribution with weights given by the observed amino-acid frequencies in the datasets. The variance in the estimation of the frequencies from a finite size sample is $\mathrm{Var}(f_i(\sigma_i)) = (p_i(\sigma_i)(1 - p_i(\sigma_i)))/N_s$ and the covariance between the frequencies of different amino-acids σ and σ' at the same site i is $\mathrm{Cov}(f_i(\sigma_i), f_i(\sigma'_i)) = -(p_i(\sigma_i)p_i(\sigma'_i))/N_s$ where N_s is the sample size and $p_i(\sigma_i)$ are the weights of the true multinomial distribution sampled. Through error propagation from these quantities we calculate the variance in the entropy of the independent sites model, to first order in $1/N_s$:

$$\mathrm{Var}(S_{indep}) = \frac{1}{N_s}\left[\sum_i \sum_{\sigma_i} p_i(\sigma_i)\log p_i(\sigma_i)^2 - S^2_{indep}\right] + O\left(\frac{1}{N_s^2}\right). \tag{24}$$

Evaluating this equation using the empirical frequencies $p = f$ assuming they are sampled from an underlying multinomial distribution, gives an estimate of the standard deviation of 0.05. We assume that the interaction terms do not change the order of magnitude of this estimation. Also the standard deviation in the averages in Eq (23) scales as $1/\sqrt{N_s}$ with N_s = 80000.

The parameter inference is affected not only by noise, but also by a systematic bias depending on the parameters of the gradient ascent described above and the initial condition that we chose to start learning from. S1 Fig shows the average entropy of 10 realizations of the learning and thermodynamic integration procedure for the ANK family and its standard deviation as error bars. If we learn the models with an increasing W window progressively we get a different profile than learning each point starting from the independent model, and above L these two profiles are more distant than the magnitude of the standard deviation, signalling a systematic bias. S1 Fig also shows that progressively learning the model results in a better parameters convergence to values that give lower entropy values.

In order to estimate how this bias is reflected in the entropy estimation we take the single-site amino-acid frequencies produced by the inferred energy function in the last Monte-Carlo phase of the learning procedure and calculate the corresponding entropy for this independent-sites model. We compute the absolute value of the difference between this estimate of the entropy and the independent-sites entropy calculated from the dataset. Again in doing this we assume that neglecting the interaction terms does not change the order of magnitude of this error. These procedure results in the errorbars shown in Figs 3 and 2, Table 1, S2 Fig.

We repeat 10 realizations of both the parameter inference procedure and the entropy estimation, and in Fig 3 we show the average entropy of these 10 numerical experiments for the ANK family where error bars are estimated as explained above to sketch the order of magnitude of the error coming from systematic bias in the parameters learning. S1 Fig shows the mean entropy of ANK as in Fig 3A with the standard deviations of the realizations entropy as error bars, to give an idea of the combined noise in the thermodynamic integration and in the gradient descent, starting from the same initial conditions and with the same update parameters (see Section). The combined noise is smaller than the entropy decrease at 33 residues, showing the decrease is real.

To further check the robustness of the entropy estimation procedure, we generate two synthetic ANK datasets, one with an independent sites model, the other with a model of two non-interacting repeats obtained as explained above, and relearn the model from the synthetic datasets. Repeating the learning and entropy estimation procedure on each on the synthetic protein families gives results that are consistent with the model used for the dataset generation. The entropy of the model learned taking an independent sites dataset does not decrease with the interaction range W and the entropy of the model learned taking a non-interacting repeats dataset does not show any drop around the repeat length.

We repeat the procedure described for the LRR and TPR repeat-proteins families reaching similar conclusions (S2 Fig).

Calculating the basins of attraction of the energy landscape

In order to characterize the ruggedness of the inferred energy landscapes and the sequence identity of the local minima, we start from all the sequences in the natural dataset as initial conditions and for each of them we perform a $T = 0$ quenched Monte-Carlo procedure. Repeating this analysis on sequences synthetically generated from E_{full} yields very similar results (see S6 Fig for ANK)

We perform this energy landscape exploration learning the parameters of the Hamiltonian in Eq 2 (refer to Section for the learning procedure), and then set $\lambda_{\text{ID}} = 0$ in the energy function because we want to investigate the shape of the energy landscape due to selection rather than the phylogenic dependence.

We scan all the possible mutations that decrease the sequence energy and then draw one of them from a uniform random distribution. The possible mutations are all single point mutations. If the same amino-acid is present in the same relative position in the two repeats we allow for double mutations that mutate those two positions to a new amino-acid, that is identical in both repeats, at the same time. We do this so that the phylogenetic biases that are still partially present in the parameters of the model do not result in spurious local minima biasing the quenching results. The Monte-Carlo procedure ends when every proposed move results in a sequence with an increased energy, and the identified sequence is a local minimum of the energy landscape.

To explore how turning on interactions makes the energy landscape more rugged, we perform the same procedure with the Hamiltonian corresponding to two intermediate interaction ranges in Fig 3A. That is Eq 2, in which J_{ij} was allowed to be non-zero only within a certain interaction range W. We picked $W = 3$ and $W = 10$.

In order to assess what is the role of the inter-repeat interactions we repeat this $T = 0$ quenched Monte-Carlo procedure on single repeats, with all the unique repeats in the natural dataset as initial condition. The learning procedure of the Hamiltonian for a single repeat is explained in Section. In this single repeat case the possible mutations are just the single point mutations.

Once we have the local minima of the energy landscape, we obtain the coarse-grained minima using the Python Scipy hierarchical clustering algorithm. In this hierarchical clustering the distance between two clusters is calculated as the average Hamming distance between all the possible pairs of sequences belonging each to one cluster. As a result we plot the clustered distance matrix, the clustering dendogram and the basin size corresponding to the distance matrix entries.

In the end we can repeat the quenching procedure described above for LRR and TPR families. The result are sketched in S3 and S4 Figs and lead to similar conclusions as for the ANK family.

Kullback-Leibler divergence

The Kullback-Leibler divergence between two families A and B is defined as $D_{\mathrm{KL}}(\mathrm{A}||\mathrm{B}) = \Sigma_{\boldsymbol{\sigma}} p_{\mathrm{A}}(\boldsymbol{\sigma}) \log_2 p_{\mathrm{B}}(\boldsymbol{\sigma})/p_{\mathrm{A}}(\boldsymbol{\sigma})$. We can substitute the sequence ensembles for ANK and TPR in the definition of the probabilities obtaining:

$$D_{\mathrm{KL}}(ANK||TPR) = \langle E_{\mathrm{TPR}} - E_{\mathrm{ANK}} \rangle_{\mathrm{ANK}} + F_{\mathrm{ANK}} - F_{\mathrm{TPR}}\,, \tag{25}$$

$$D_{\mathrm{KL}}(TPR||ANK) = \langle E_{\mathrm{ANK}} - E_{\mathrm{TPR}} \rangle_{\mathrm{TPR}} + F_{\mathrm{TPR}} - F_{\mathrm{ANK}}\,, \tag{26}$$

where the notation $\langle\rangle_{\mathrm{ANK}}$ means that the average is calculated over sequences drawn from the ANK ensemble: $P(\boldsymbol{\sigma})_{\mathrm{ANK}} = (1/Z_{\mathrm{ANK}})e^{-E(\boldsymbol{\sigma})_{\mathrm{ANK}}}$. Therefore $\langle E_{\mathrm{TPR}} \rangle_{\mathrm{ANK}}$ is the average TPR energy function evaluated, via the structural alignment between the two families, on 80000 sequences generated through a Monte Carlo sampling of the ANK model 2 (and analogously for $\langle E_{\mathrm{ANK}} \rangle_{\mathrm{TPR}}$). The terms F_{ANK} and F_{TPR} are calculated in the same way as when estimating the entropy through Eqs 21 and 22, as explained in Section.

For the control against a random polypeptide of length L we use $D_{\mathrm{KL}}(\mathrm{FAM}||\mathrm{rand}) = \log \Lambda - S(\mathrm{FAM})$, where $\Lambda = 21^L$ is the total number of possible sequences of length L.

Supporting information

S1 Fig. Reproducibility of entropy estimation. Entropy as a function of the maximum linear interaction range W along the sequence. Green curve: entropy of the ANK family with error bars calculated as standard deviations over 10 model learning realizations, where models are learned by incrementally adding more interaction terms as W is increased, taking the model learned at $W - 1$ as initial condition. This plot is the same as in Fig 3A but with the different error bar estimates, showing that our results are robust to the details of error estimation. Red curve: entropy obtained after *de novo* learning for each W, starting from a non-interacting model as initial condition. With those initial conditions the learning gets stuck, leading to systematically overestimating the entropy and missing the second entropy drop at $W = L - 1$. See Section for details of the learning and entropy estimation procedure.
(TIFF)

S2 Fig. Range dependence of entropy in LRR and TPR families. Entropy of the LRR (A) and TPR (B) family as a function of the maximum interaction distance W along the sequence. The entropy of the model decreases as a more interactions are added and they constrain the space of possible sequences. As with ANK, the entropy first drops, plateaus, then drops again at the distance corresponding to homologous positions along the two repeats ($W = L - 1 = 23$ for LRR, and 33 for TPR, dashed line). This second drop indicates that there is a typical distance along the sequence, corresponding to the repeat length, where interactions due to structural properties constrain the sequence ensemble. The error bars are estimated approximately from errors in learning (see Section). Entropies are averaged over 5 realizations of the learning and entropy estimation procedure.
(TIFF)

S3 Fig. Analysis of local energy minima for pairs of consecutive repeats of LRR. Energy minima were obtained by zero-temperature dynamics. Sequences falling into a given minimum with these dynamics define its basin of attraction. A, bottom) rank-frequency plot of the sizes of the basins of attraction. A, top) energy minimum of each basin. Gray line shows the energy of the consensus sequence B) Pairwise Hamming distances between energy minima, organised by hierarchical clustering. The panel right above the matrix shows the the size of the

basins relative to the minima corresponding to the entries of the distance matrix. C and D) Same analysis as A) and B), but for single LRR repeats.
(TIFF)

S4 Fig. Analysis of local energy minima for pairs of consecutive repeats of TPR. Energy minima were obtained by zero-temperature dynamics. Sequences falling into a given minimum with these dynamics define its basin of attraction. A, bottom) rank-frequency plot of the sizes of the basins of attraction. A, top) energy minimum of each basin. Gray line shows the energy of the consensus sequence B) Pairwise Hamming distances between energy minima, organised by hierarchical clustering. The panel right above the matrix shows the the size of the basins relative to the minima corresponding to the entries of the distance matrix. C and D) Same analysis as A) and B), but for single TPR repeats.
(TIFF)

S5 Fig. Interactions within repeats increase the ruggedness of the energy landscape. Local minima were obtained by performing a zero-temperature Monte-Carlo simulation with the energy function in Eq (2) with non-zero J_{ij} within linear interaction range W, starting from initial conditions corresponding to naturally occurring sequences of pairs of consecutive ANK repeats, for $W = 3$ (A and B) and $W = 10$ (C and D). See Fig 5 for the full model ($W = 2L$). A and C, bottom: Rank-frequency plot of basin sizes, where basins are defined by the set of sequences falling into a particular minimum. A and C, top: energy of local minima *vs* the size-rank of their basin. Gray line indicates the energy of the consensus sequence, for comparison. B and D: Pairwise distance between the minima with the largest basins (comprising 90% of natural sequences), organised by hierarchical clustering. The panel right above the matrix shows the size of the basins relative to the minima corresponding to the entries of the distance matrix. The block structure starts emerging as interactions are turned on (D versus B).
(TIFF)

S6 Fig. Analysis of local energy minima from generated pairs of consecutive repeats of ANK. Energy minima were obtained by zero-temperature dynamics starting from sequences generated *in silico* from E_{full}. Sequences falling into a given minimum with these dynamics define its basin of attraction. A, bottom) rank-frequency plot of the sizes of the basins of attraction. A, top) energy minimum of each basin. Gray line shows the energy of the consensus sequence B) Pairwise Hamming distances between energy minima, organised by hierarchical clustering. The panel right above the matrix shows the size of the basins relative to the minima corresponding to the entries of the distance matrix. C and D) Same analysis as A) and B), but for single ANK repeats.
(TIFF)

Author Contributions

Conceptualization: Diego U. Ferreiro, Aleksandra M. Walczak, Thierry Mora.

Data curation: Ezequiel A. Galpern, Rocio Espada, Diego U. Ferreiro.

Formal analysis: Jacopo Marchi, Ezequiel A. Galpern, Aleksandra M. Walczak, Thierry Mora.

Funding acquisition: Diego U. Ferreiro, Aleksandra M. Walczak.

Investigation: Jacopo Marchi, Diego U. Ferreiro, Aleksandra M. Walczak, Thierry Mora.

Methodology: Jacopo Marchi, Ezequiel A. Galpern, Thierry Mora.

Project administration: Aleksandra M. Walczak, Thierry Mora.

Supervision: Aleksandra M. Walczak, Thierry Mora.

Validation: Jacopo Marchi.

Visualization: Jacopo Marchi.

Writing – original draft: Jacopo Marchi, Aleksandra M. Walczak, Thierry Mora.

Writing – review & editing: Diego U. Ferreiro, Aleksandra M. Walczak, Thierry Mora.

References

1. Dryden DTF, Thomson AR, White JH. How much of protein sequence space has been explored by life on Earth? Journal of the Royal Society InterfaceRoyal Society Interface. 2008; 5(April):953–956. https://doi.org/10.1098/rsif.2008.0085
2. Shakhnovich EI. Protein design: a perspective from simple tractable models. Current Biology. 1998; 3:45–58.
3. Salisbury FB. Natural Selection and the Complexity of the Gene. Nature. 1969; 244:342–343. https://doi.org/10.1038/224342a0
4. Mandecki W. The game of chess and searches in protein sequence space. Biotopic. 1998; 16(May): 200–202.
5. Dokholyan NV, Shakhnovich B, Shakhnovich EI. Expanding protein universe and its origin from the biological Big Bang. PNAS. 2002; 99:14132–14136. https://doi.org/10.1073/pnas.202497999 PMID: 12384571
6. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, et al. The Pfam protein families database. Nucleic acids research. 2004; 32(suppl 1):D138–D141. https://doi.org/10.1093/nar/gkh121 PMID: 14681378
7. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic acids research. 2013; p. gkt1223.
8. Neher E. How frequent are correlated changes in families of protein sequences? Proceedings of the National Academy of Sciences. 1994; 91(1):98–102. https://doi.org/10.1073/pnas.91.1.98
9. Morcos F, Hwa T, Onuchic JN, Weigt M. Direct coupling analysis for protein contact prediction. Methods Mol Biol. 2014; 1137:55–70. https://doi.org/10.1007/978-1-4939-0366-5_5 PMID: 24573474
10. Szurmant H, Weigt M. Inter-residue, inter-protein and inter-family coevolution: bridging bridging the scales. Current Opinion in Structural Biology. 2018; 50:26–32. https://doi.org/10.1016/j.sbi.2017.10.014 PMID: 29101847
11. Tubiana J, Cocco S, Monasson R. Learning protein constitutive motifs from sequence data. eLife. 2019; 8:e39397:1–61. https://doi.org/10.7554/eLife.39397 PMID: 30857591
12. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein–protein interaction by message passing. Proceedings of the National Academy of Sciences. 2009; 106(1):67–72. https://doi.org/10.1073/pnas.0805923106
13. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences. 2011; 108(49):E1293–E1301. https://doi.org/10.1073/pnas.1111471108
14. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Physical Review E. 2013; 87(1):012707. https://doi.org/10.1103/PhysRevE.87.012707
15. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. Cell. 2012; 149(7):1607–1621. https://doi.org/10.1016/j.cell.2012.04.012 PMID: 22579045
16. Espada R, Parra RG, Mora T, Walczak AM, Ferreiro DU. Capturing coevolutionary signals inrepeat proteins. BMC bioinformatics. 2015; 16(1):207. https://doi.org/10.1186/s12859-015-0648-3 PMID: 26134293
17. Espada R, Parra RG, Mora T, Walczak AM, Ferreiro DU. Inferring repeat-protein energetics from evolutionary information. PLoS computational biology. 2017; p. 1–16.
18. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H. High-resolution protein complexes from integrating genomic information with molecular simulation. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106(52):22124–9. https://doi.org/10.1073/pnas.0912100106 PMID: 20018738

19. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PloS one. 2011; 6(12):e28766. https://doi.org/10.1371/journal.pone.0028766 PMID: 22163331

20. Contini A, Tiana G. A many-body term improves the accuracy of effective potentials based on protein coevolutionary data. The Journal of Chemical Physics. 2015; 143(2):025103. https://doi.org/10.1063/1.4926665 PMID: 26178131

21. Haldane A, Flynn WF, He P, Vijayan RSK, Levy RM. Structural propensities of kinase family proteins from a Potts model of residue co-variation. Protein Science; 25(8):1378–1384. https://doi.org/10.1002/pro.2954 PMID: 27241634

22. Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. Molecular biology and evolution. 2015;. https://doi.org/10.1093/molbev/msv211 PMID: 26446903

23. Maynard Smith J. Natural Selection and the Concept of a Protein Space. Nature. 1970; 225:563–564. https://doi.org/10.1038/225563a0

24. Weinreich DM, Delaney NF, Depristo MA, Hartl DL. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science (New York, NY). 2006; 312(5770):111–114. https://doi.org/10.1126/science.1123539

25. Li J, Mahajan A, Tsai Md, Li J, Mahajan A, Tsai Md. Current Topics / Perspectives Ankyrin Repeat: A Unique Motif Mediating Protein − Protein Interactions Ankyrin Repeat: A Unique Motif Mediating Protein—Protein Interactions †. Biochemistry. 2006; 45(November):15168–15178.

26. Tian P, Best RB. How Many Protein Sequences Fold to a Given Structure? A Coevolutionary Analysis. Biophysj. 2017; 113(8):1719–1730. https://doi.org/10.1016/j.bpj.2017.08.039

27. Barton JP, Chakraborty AK, Cocco S, Jacquin H, Monasson R. On the Entropy of Protein Families. Journal of Statistical Physics. 2016; 162(5):1267–1293. https://doi.org/10.1007/s10955-015-1441-4

28. Jaynes ET. Information Theory and Statistical Mechanics. Phys Rev. 1957; 106:620–630. https://doi.org/10.1103/PhysRev.106.620

29. Tian P, Louis JM, Baber JL, Aniana A, Best RB. Co-Evolutionary Fitness Landscapes for Sequence Design. Angew Chemie—Int Ed. 2018; 57(20):5674–5678. https://doi.org/10.1002/anie.201713220

30. Shakhnovich EI, Gutin AM. A new approach to the design of stable proteins. Protein Eng. 1993; 6(8):793–800. https://doi.org/10.1093/protein/6.8.793 PMID: 8309926

31. Shakhnovich EI, Gutin AM. Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci. 1993; 90(15):7195–7199. https://doi.org/10.1073/pnas.90.15.7195 PMID: 8346235

32. Dokholyan NV, Shakhnovich EI. Understanding Hierarchical Protein Evolution from First Principles. Journal of Molecular Biology. 2001; 312:289–307. https://doi.org/10.1006/jmbi.2001.4949 PMID: 11545603

33. Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. 2014; 111(34):12408–12413.

34. Kimura M. On the Probability of Fixation of Mutant Genes in a Population. Genetics. 1962; 47:713–719. PMID: 14456043

35. Berg J, Willmann S, Lässig M. Adaptive evolution of transcription factor binding sites. BMC evolutionary biology. 2004; 4:42. https://doi.org/10.1186/1471-2148-4-42 PMID: 15511291

36. Mezard M, Parisi G, Virasoro M. Spin Glass Theory and Beyond. WORLD SCIENTIFIC; 1986. Available from: https://www.worldscientific.com/doi/abs/10.1142/0271.

37. Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. Proceedings of the National Academy of Sciences. 1987; 84(21):7524–7528. https://doi.org/10.1073/pnas.84.21.7524

38. Frauenfelder H, Sligar SG, Wolynes PG. Proteins. Science. 1991; 254:1598–1603.

39. Facco E, Pagnani A, Russo ET, Laio A. The intrinsic dimension of protein sequence evolution. PLoS Comput Biol. 2019; 15(4):e1006767. https://doi.org/10.1371/journal.pcbi.1006767 PMID: 30958823

40. Boersma YL, Plu A. DARPins and other repeat protein scaffolds: advances in engineering and applications. Current opinion in biotechnology. 2011; 22:849–857. https://doi.org/10.1016/j.copbio.2011.06.004 PMID: 21715155

41. Tripp KW, Barrick D. Rerouting the Folding Pathway of the Notch Ankyrin Domain by Reshaping the Energy Landscape. Journal of the American Chemical Society. 2008;(19):5681–5688. https://doi.org/10.1021/ja0763201 PMID: 18396879

42. Barrick D, Ferreiro DU, Komives EA. Folding landscapes of ankyrin repeat proteins: experiments meet theory. Current Opinion in structural biology. 2008; 18:27–43. https://doi.org/10.1016/j.sbi.2007.12.004 PMID: 18243686

43. Consortium U, et al. UniProt: the universal protein knowledgebase. Nucleic acids research. 2017; 45(D1):D158–D169. https://doi.org/10.1093/nar/gkw1099

44. Li W, Jaroszewski L, Godzik A. Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics. 2002; 18(1):77–82. https://doi.org/10.1093/bioinformatics/18.1.77 PMID: 11836214

45. Mosavi LK, Minor DL, Peng Zy. Consensus-derived structural determinants of the ankyrin repeat motif. Proceedings of the National Academy of Sciences. 2002; 99(25):16029–16034. https://doi.org/10.1073/pnas.252537899

46. 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. Journal of molecular biology. 2003; 332(2):489–503. https://doi.org/10.1016/s0022-2836(03)00896-9 PMID: 12948497

47. Frankel D, Smit B. Understanding Molecular Simulation: From Algorithms to Applications. Academic Press; 2007.