PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Inferring repeat-protein energetics from evolutionary information

Rocío Espada[1], R. Gonzalo Parra[2], Thierry Mora[3], Aleksandra M. Walczak[4], Diego U. Ferreiro[1]*

1 Protein Physiology Lab, Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Química Biológica. Buenos Aires, Argentina. / CONICET - Universidad de Buenos Aires. Instituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales (IQUIBICEN). Buenos Aires, Argentina, **2** Quantitative and Computational Biology Group, Max Planck Institute for Biophysical Chemistry, Goettingen, Germany, **3** Laboratoire de physique statistique, Ecole Normale Supérieure, CNRS and UPMC, 75005 Paris, France, **4** CNRS and Laboratoire de Physique Théorique, Ecole Normale Supérieure, Paris, France

* ferreiro@qb.fcen.uba.ar

Citation: Espada R, Parra RG, Mora T, Walczak AM, Ferreiro DU (2017) Inferring repeat-protein energetics from evolutionary information. PLoS Comput Biol 13(6): e1005584. https://doi.org/10.1371/journal.pcbi.1005584

Editor: Alexandre V Morozov, Rutgers University, UNITED STATES

Received: February 7, 2017

Accepted: May 21, 2017

Published: June 15, 2017

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

Funding: This work was supported by AGENCIA NACIONAL DE PROMOCIÓN CIENTÍFICA Y TECNOLÓGICA - http://www.agencia.mincyt.gob.ar/frontend/agencia/fondo/foncyt grants: PICT 2012-0164, ERCStG n. 306312, and ECOS Sud - MINCyT n° A14E04. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Natural protein sequences contain a record of their history. A common constraint in a given protein family is the ability to fold to specific structures, and it has been shown possible to infer the main native ensemble by analyzing covariations in extant sequences. Still, many natural proteins that fold into the same structural topology show different stabilization energies, and these are often related to their physiological behavior. We propose a description for the energetic variation given by sequence modifications in repeat proteins, systems for which the overall problem is simplified by their inherent symmetry. We explicitly account for single amino acid and pair-wise interactions and treat higher order correlations with a single term. We show that the resulting evolutionary field can be interpreted with structural detail. We trace the variations in the energetic scores of natural proteins and relate them to their experimental characterization. The resulting energetic evolutionary field allows the prediction of the folding free energy change for several mutants, and can be used to generate synthetic sequences that are statistically indistinguishable from the natural counterparts.

Author summary

Unlike most natural proteins that are made with apparently random strings of amino acids, repeat-proteins are formed with tandem stretches of similar elements. The statistical description for these occurrences can be captured with a simple energetic model that accounts for evolutionary mechanism that gave rise to these proteins. The resulting energetic model can be used to infer folding stability and can generate sequences that are indistinguishable from the natural ones.

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

Introduction

Repeat proteins are composed of tandem repetitions of similar structural motifs of about 20 to 40 amino acids. Under appropriate conditions, these polymers fold into elongated, non-globular structures (Fig 1). It is apparent that the overall architecture is stabilized mainly by short range interactions, in contrast to most globular protein domains that usually adopt very intricate topologies [1]. In their natural context, repeat proteins are frequently found mediating protein-protein interactions, with a specificity rivaling that of antibodies [2–4]. Given their structural simplicity and potential technological applications, repeat-proteins are a prime target for protein design, with very successful examples for a variety of topologies [5–7]. Most of the current design strategies target the creation of rigid native structures with desired folds that, although beautiful, often lose biological functionality [8]. It is becoming clear that the population of 'excited states' is crucial for protein function [9], and thus tackling energetic inhomogeneities in protein structures may be crucial for understanding how biological activities emerge [10]. The challenge thus relies in finding an appropriate description for the 'energy' of each system, a daunting task for large molecular objects such as natural proteins.

In principle, the natural variations observed for proteins of the same family must contain information about the sequence-structure mapping. A simple model that just takes into account the frequency of each amino acid in each position is insufficient to capture collective effects, yet, for some architectures it is surprisingly good for the synthesis of non-natural repeat-proteins by 'consensus' design [11–14]. It is apparent that in the case of repeat proteins the local signals play inordinately large roles in the energy distribution, just as expected from their topology [15] and hence, small heterogeneities can be propagated from the local repeat units to higher orders affecting the overall structure and dynamics [16, 17]. Thus, collective effects may be approximated as small perturbations to local potentials, simplifying the energetic description of complex natural systems [18].

In the last years new methods to analyze correlated mutations across a family of proteins have arisen (mfDCA [19], plmDCA [20, 21], Gremlin [22] to name a few). The main hypothesis behind these methods is that biochemical changes produced by a point mutation should be compensated by other mutations (along evolutionary timescales) to maintain protein viability or function. These methods can also be used to disentangle relevant direct correlations from indirect ones. They are very successful at predicting spatial contacts and interactions for many protein topologies [23–27]. Nevertheless, these methods do not take into account the chemical nature of the amino acids, which can be codifying inhomogeneities in the energetic distribution that are crucial for the activity of repeat-proteins [28, 29]. On this basis, different approaches have been proposed recently to include chemical details in the correlation analyses [30], trying to predict folding stability [31], conformational heterogeneity [23, 32, 33], mutational effect in the interaction in two-component signaling proteins [27] or the global effect on

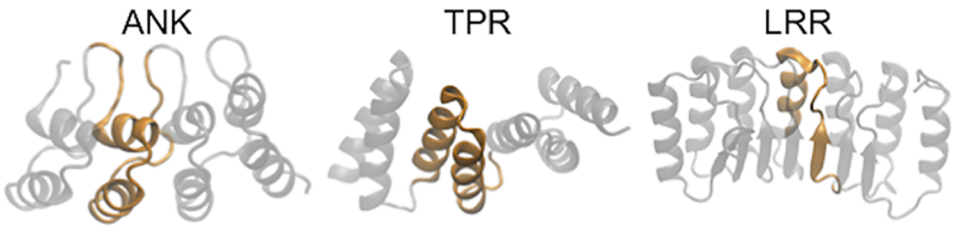

Fig 1. Repeat proteins are elongated objects with internal symmetry. Representative structures of members of the repeat proteins families studied. On left, ankyrin repeat (PDB id:1N0R [38]), center tetratricopeptide-like repeats (PDB id:1NA0 [11]) and right leucine-rich repeats (PDB id:4IM6). The defined repeated unit is highlighted in orange.

https://doi.org/10.1371/journal.pcbi.1005584.g001

antibiotic resistance from sequences of β-lactamases [34, 35]. As many other tools, these were optimized to perform well on globular proteins, and their application to repeat proteins is not straightforward. Besides the point-mutation mechanism, repeat proteins are believed to evolve via duplication and rearrangement of repeats [36], resulting in an inherent symmetry which usually confounds sequence analyses [17]. Making use of this symmetry, we have previously proposed a specific version of mfDCA and plmDCA for repeat proteins [37]. In this work we develop an alternative 'evolutionary field', able to disentangle biases generated by the repetitive nature of these proteins and which explicitly includes the information of the amino acids that compose a protein. We show that it is able to reflect biochemical properties of the analyzed proteins. We take advantage of the elongated and repetitive structure of these proteins (Fig 1) to extract as much information as possible from the data, and apply the general ideas on three specific families, ankyrin repeats (ANK), leucine-rich repeats (LRR) and tetratricopeptide-like repeats (TPR).

Evolutionary energy for repeat proteins

To study the co-occurrence of mutations in a sequence alignment of a particular protein family, [39] proposed a Hamiltonian or energy expression which resembles a Potts model:

$$E(\vec{s}) = -\left[\sum_{i=1}^{L} h_i(a_i) + \sum_{i=1}^{L}\sum_{j=i}^{L} J_{ij}(a_i, b_j)\right] \tag{1}$$

where the set of $\{h_i(a_i)\}$ parameters, one for each amino acid in each position, accounts for a local propensity of having a specific residue on a particular site of the protein, and the set of $\{J_{ij}(a_i, b_j)\}$ indicates the strength of the 'evolutionary' interaction between each possible amino acid in every pair of positions along the protein. There are $q = 21$ possible values of a_i and b_j, one for each amino acid and one for the gaps included on the multiple sequence alignments. This expression is evaluated on a particular sequence on an alignment, and the summations go over the L columns of the alignment. A sequence is more favorable or more energetic if it gets lower values of $E(\vec{s})$. It can be expected that the population of sequences follows a Boltzmann distribution $P(\vec{s}) = \frac{1}{Z} e^{-E(\vec{s})}$ [40]. The parameters are thus fitted to reproduce the frequencies of occurrence of each amino acid in each position ($f_i(a_i)$) and the joint frequencies of amino acids ($f_{ij}(a_i, b_j)$) in an alignment of natural sequences used as input:

$$f_i(a_i) = \sum_{a_k, k\neq i} P(\vec{s}) \tag{2}$$

$$f_{ij}(a_i, b_j) = \sum_{a_k, k\neq i,j} P(\vec{s}) \tag{3}$$

Nevertheless, for repeat proteins there is another feature we want to capture with an evolutionary energy: the high identity of amino acids constituting consecutive repeats, arisen by the repetitiveness of these families and probably a signature of their evolutionary mechanisms (Fig 2).

Therefore, we propose the following model for repeat proteins:

$$E(\vec{s}) = -\left[\sum_{i=1}^{L} h_i(a_i) + \sum_{i=1}^{L}\sum_{j=i}^{L} J_{ij}(a_i, b_j) - \lambda_{Id}(\vec{s})\right] \tag{4}$$

This expression is designed to be applied in sequences constituted by two repeats. λ_{Id} is a parameter that aims at reproducing the probabilities of the percentage of identity (%Id) between consecutive repeats in natural proteins (p_{id}). Basically, it accounts for higher order

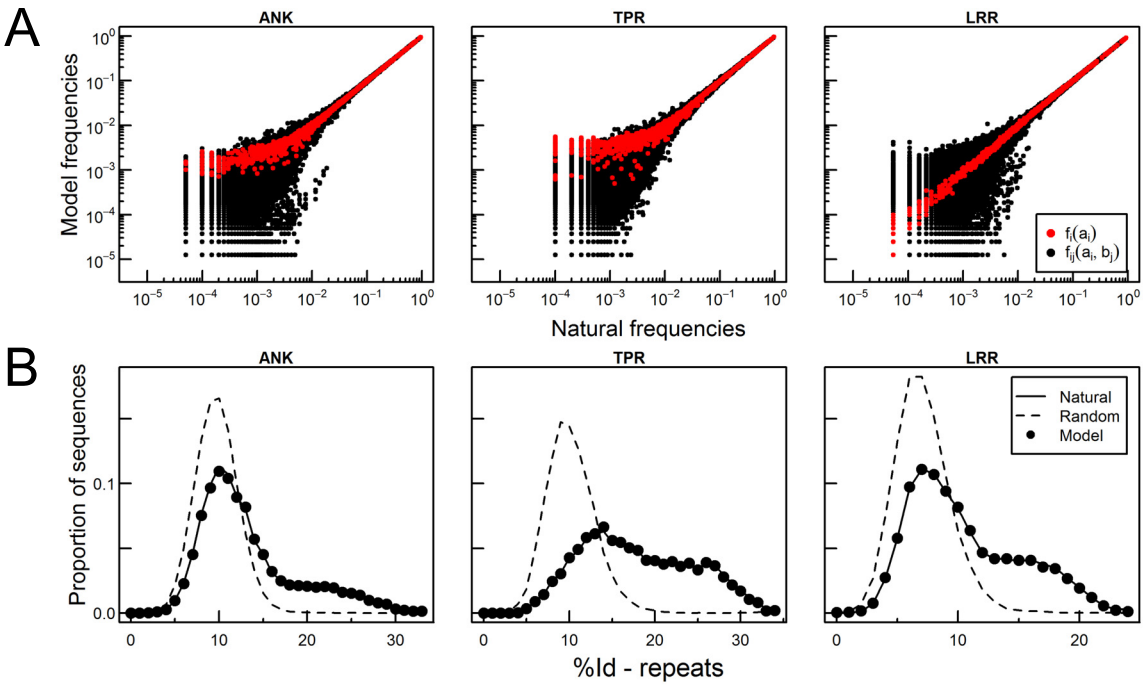

Fig 2. The proposed model fits the frequencies of amino acids and natural repeat identities p_{id}. On A, we compare marginal frequencies $f_i(a_i)$ (red) and joint frequencies $f_{ij}(a_i, b_j)$ (black) on the natural ensemble of sequences (x-axis) and on the set of sequences generated by the model (y-axis). On B, we calculate the distribution of identity between repeats p_{id} for consecutive repeats (solid line), and for natural repeats which are not consecutive, i.e. they are not next to each other in the primary structure (dot lines). Consecutive repeats present a population with high identity between repeats that any pairs of repeats do not show. We compare the distribution produced by the model p_{id}^{model} (dots).

https://doi.org/10.1371/journal.pcbi.1005584.g002

correlations not captured by the pairwise terms. For a given sequence we calculate the %Id of the adjacent repeats and sum the parameter λ_{Id} corresponding to that %Id value. When the correct parameters are obtained, this equation can be used to produce an ensemble of sequences consistent with the constraints ($f_i(a_i), f_{ij}(a_i, b_j)$ and p_{id}). We work with pairs of repeats as it is the minimum unit that includes the interaction between repeats and the possibility of measure sequence identity between consecutive repeats. In the following section we will show the convergence of the method and the relevant information that can be obtained from it. For further details about the procedure to assign values to the parameters, please refer to Methods section.

Results

Evolutionary energy reproduces ensembles of sequences with natural frequencies and repeat protein characteristics

We construct an alignment of pairs of repeats for each family: ANK (PFAM id PF00023, and final alignment of 20513 sequences of L = 66 residues each), TPR (PFAM id PF00515, and final alignment of 10020 sequences of L = 68 residues each) and LRR (PFAM id PF13516, and final alignment of 18839 sequences of L = 48 residues each). See Methods for further details of construction. We measure $f_i(a_i), f_{ij}(a_i, b_j)$ and p_{id}. Using a gradient descent procedure we obtain a set of parameters in eq 4 which are able to reproduce $f_i(a_i), f_{ij}(a_i, b_j)$ and p_{id}. In principle, the number of parameters is large: Lq h_i parameters, $\frac{(Lq)^2 - Lq}{2}$ J_{ij} parameters and $\frac{L}{2} + 1$ λ_{Id}. For example, for pairs of ANK repeats this means 1386 h_i, 959805 J_{ij} and 34 λ_{Id}. To reduce the

number of free parameters to fit we use a L_1-regularization which fixes to zero those parameters which do not contribute significantly to fit the frequencies. This regularization allows us to set to exactly zero between 85 and 91% of the J_{ij} parameters which, when they are free to vary, only reach small values (S3 Fig). We bound the maximum error permitted in the frequency estimations to 0.02. Refer to Methods for more details.

In the three families studied, the parameters obtained allow us to generate ensembles of sequences which reproduce natural $f_i(a_i)$, $f_{ij}(a_i, b_j)$ and p_{id} (Fig 2A). Notice that most frequencies are fitted with an error of an order of magnitude lower than the maximum bound imposed (S2 Fig).

The p_{id} distributions are also very well reproduced (Fig 2B). Not only the general shape, but also the populated long tail for highly similar repeats. It is not possible to obtain the same distribution only by fitting amino acid frequencies $f_i(a_i)$ and $f_{ij}(a_i, b_j)$, it is mandatory to explicitly include the p_{id} by including the parameters λ_{Id} (S1 Fig), suggesting that higher order correlations must be accounted for describing these systems.

Evolutionary energy distinguishes between proteins on a given family and other polypeptides

Once the set of parameters $\{h_i(a_i), J_{ij}(a_i, b_j), \lambda_{Id}\}$ is obtained, it can be used to score any sequence of L amino acids via eq 4. In this section we test if this measure is capable of distinguishing polypeptides that fold in a three dimensional structure similar to members of the repeat protein family from those that do not.

We calculate the distribution of energies of different sets of sequences (Fig 3). The ensembles of natural sequences of each protein family used to learn the parameters have a unimodal distribution of energies centered around -100 (Fig 3, red lines). These distributions are clearly differentiated from the energies of random chains of residues (Fig 3, yellow lines), which constitute a basic negative control for our model.

For a positive control we evaluate designed proteins which have been experimentally synthesized. For the ANK family, we consider the library of repeat sequences built by Plückthun's laboratory [13] (green lines, Fig 3A). This library was constructed by fixing on each repeat

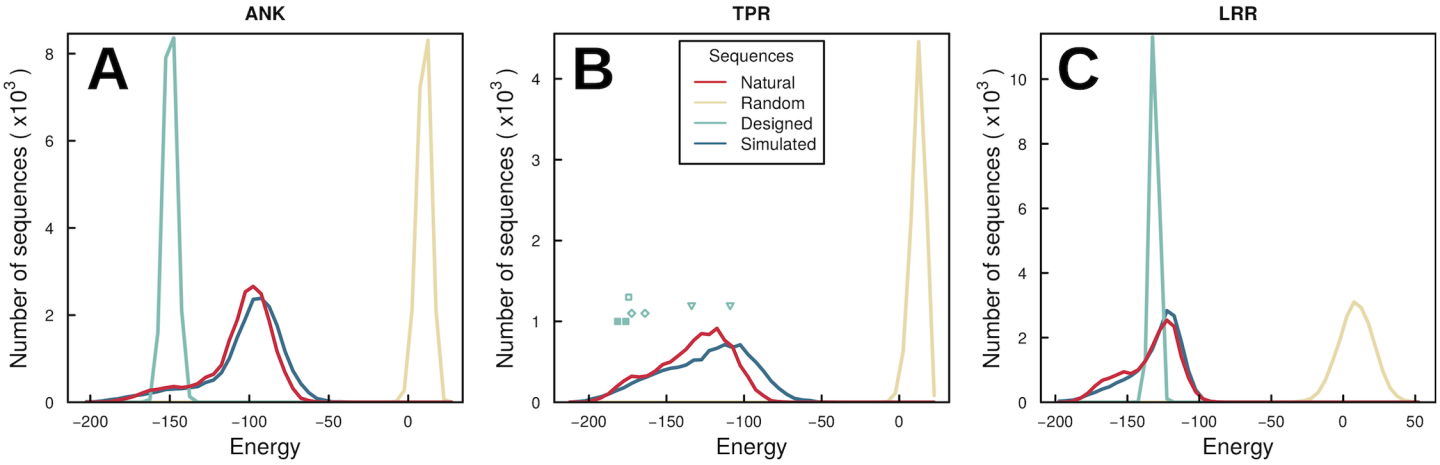

Fig 3. Energy score distribution for different ensembles of sequences for ANK (A panel), TPR (B) and LRR (C). Red lines, natural sequences used to train the model on eq 4. Blue lines, sequences simulated by Monte Carlo under expression 4. In the three families, it overlaps with natural sequences, suggesting that simulated sequences imitate the natural ensemble. Yellow lines, strings of random amino acids used as negative control. They show that the energy distinguishes between polypeptides belonging to a protein family and other strings of amino acids. Green lines, squares, diamonds and triangles, energies for designed proteins.

https://doi.org/10.1371/journal.pcbi.1005584.g003

26 positions out of 33 to the most frequent residue in the multiple sequence alignment. This resulted in a set of sequences that have small variations with respect to the ANK consensus (the sequence with the most frequent amino acid in each position). In our expression, they score a very low energy distribution, overlapping with the most negative tail of the distribution of natural sequences. It is notable that consensus designed ANK have been shown experimentally to be extremely stable. For the TPR family, consensus designed was done by Regan's laboratory [11, 12]. All pairs of repeats synthesized have the same amino acid sequence, and it's energy score is indicated by a green full square in Fig 3B. Again, the designed sequence matches values at the most left side of the energy distribution of natural sequences, and coincidentally reports high folding stability. From it, other variants with few point mutations to improve binding to a specific ligand have been synthesized. As shown in empty green squares [41] and diamonds [42] in Fig 3B, they have higher energy, but still in the left most side of natural sequences distribution. Recently, a different design strategy was done [43]. Based on a non-repetitive protein, but similar to TPR fold, they put togheter various repetitions of the fold, using TPR loops to link them. They obtained a three-repeats protein whose pair of repeats energy are represented on triangles on Fig 3B. This time, they match natural sequences distribution in higher values.

Finally, for the LRR family we contrast with the library of proteins designed by Plückthun's group based on the consensus sequence [14]. The repetition they considered has 57 amino acids, which includes two types of repeats, one of 28 residues and the other one of 29. As the repeat we are using for LRR is 24 residues long, we aligned both definitions and evaluated the library removing the amino acids not matching our definition. Again, their scores form a narrow distribution, but this time it is not placed on the most favorable side of the natural sequences distribution (Fig 3C). Coincidentally, selected species studied do not show such a high folding stability as the ANK library did.

With these parameters, we are able to generate an ensemble of sequences which are in agreement with the constraints used, via a Monte Carlo simulation (see Methods). The distribution of energies of these simulated sequences matches the natural sequences energies distribution with remarkable accuracy. Moreover, we randomly choose 100 sequences from the natural ensemble and 100 sequences from the simulated one, perform a Smith-Waterman pairwise alignment all against all, calculate the pair similarity using BLOSUM62 matrix and used it as a distance method to plot a dendogram of the sequences (S4 Fig). Both species appear interspersed, showing that it is not possible to distinguish a natural sequence from a constructed one. Also, we tested *familiarity* to the ANK family as defined in [44] and found overlapping distributions for both species (S5 Fig). Therefore, simulated sequences represent possible variants to natural repeats. The wide distribution of natural proteins suggests that it should be possible to engineer sequences with more variable repeats, more dissimilar among neighbors and to the consensus than the ones published up to date.

Low evolutionary energy sequences have similar repeats

Are there any invariant properties shared by low energy sequences? Given that repeat-proteins may evolve by other mechanisms besides point substitutions, we analyze if low energy sequences are constituted by highly similar repeats and if they are close to consensus sequences.

On Fig 4A we show the relation between the %Id between the repeats and the energy of the sequence. It is evident that low energy sequences are constructed by pairs of highly similar repeats. This could be a transitive effect: if low energy sequences are very similar to the consensus sequence, and the consensus sequence is formed by two identical repeats, we would be

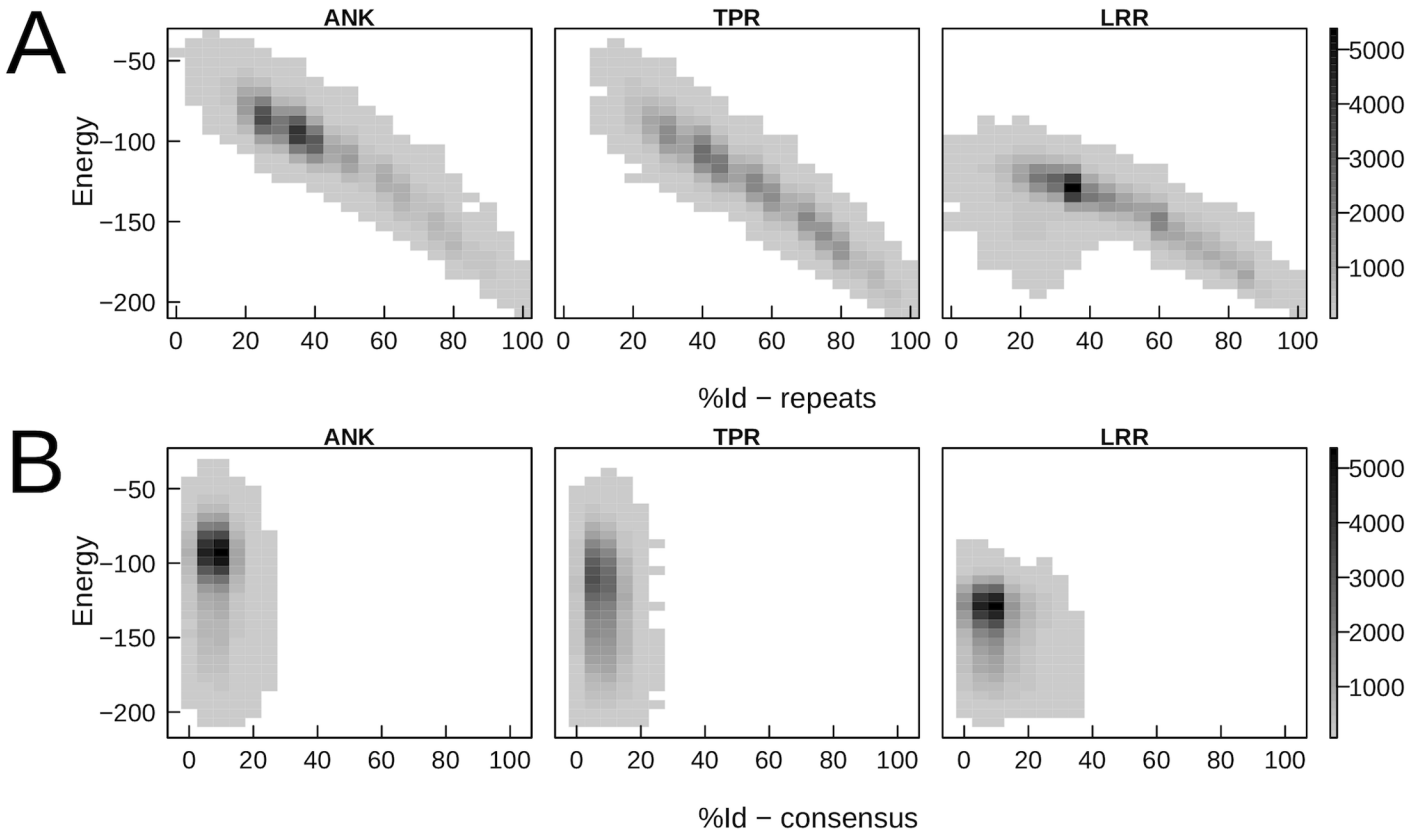

Fig 4. Most favorable simulated sequences have very similar repeats, yet they are different to the consensus repeat. On A, we plot the energy vs. the identity between the repeats that constitute the sequence. Even though the deviation is large, most stable sequences tend to have more similar repeats. On B, we plot the energies of simulated sequences vs the identity to the consensus of the family. In all cases, the identity to the consensus is small and uncorrelated to the energy, indicating that sequences which differ significantly from the consensus can be stable variants of the family.

https://doi.org/10.1371/journal.pcbi.1005584.g004

seeing that more similarity between repeats causes lower energies. We can see that it is not the case (Fig 4B). We plot the %Id to the consensus against the energy of each sequence. The consensus was calculated with the most frequent amino acid in each position on sequences used as input. We can see that there is no evident correlation between the energy and the similarity to the consensus. Thus, low energy sequences that differ from the consensus one may be constructed. Also, there are no sequences which get a high %Id to the consensus. We conclude that there are different repeats which have low energies within a protein family, and not only the consensus sequence.

Evolutionary energy and folding stability change upon point mutations

Consensus designed ANK proteins are very stable upon chemical and thermal denaturation [13], and, as shown in Fig 3 also score a very low evolutionary energy according to eq 4. Can we quantify the relationship between the stability and the evolutionary energy?

A potential test can be performed by comparing to experiments in which the effect of point mutations was evaluated. These incorporate one, two or three point mutations in natural proteins, and characterize the unfolding free energy ΔG of the wildtype and the mutated variant. A higher ΔG reports a more stable protein. We compare the change in the ΔG between the mutated and the wildtype protein ($\Delta\Delta G$), and the difference of energy for their sequences according to eq 4.

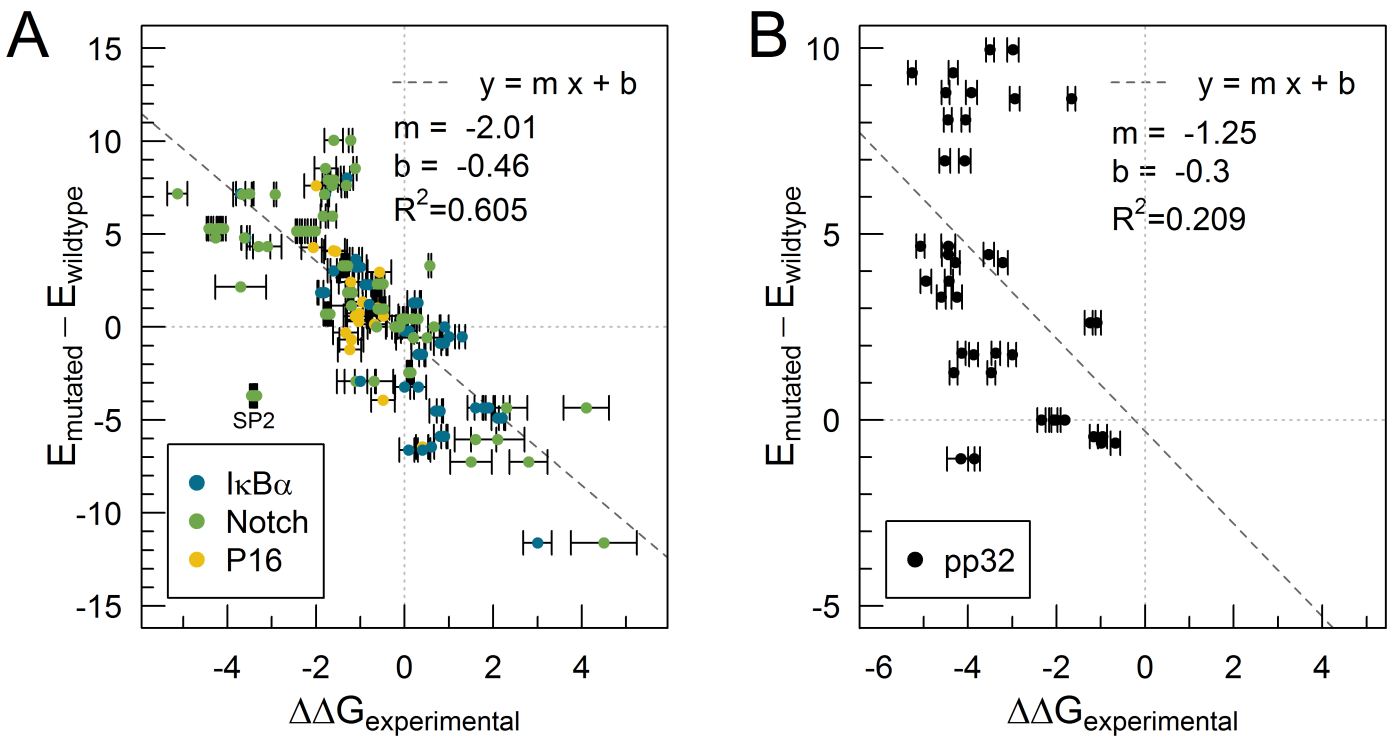

Fig 5. Variation of energy score as a predictor of the folding stability upon point mutations. We compare difference in unfolding ΔG between a wildtype protein and a mutated variant (x-axis) and the change in energy according to Eq 4. Error bars indicate the experimental standard deviation. On A, for proteins belonging to ANK family, and on B for LRR.

https://doi.org/10.1371/journal.pcbi.1005584.g005

Although the energy expression is learned for pairs of repeats, we can easily extend it to an array of repeats making use of the elongated structure of repeat proteins in which only adjacent repeats interact. From our expression we have parameters assigned to intra-repeat positions (h_i with $i = 1 \ldots \frac{L}{2}$ and J_{ij} with $i, j = 1 \ldots \frac{L}{2}$), and inter-repeat interactions (J_{ij} with $i = 1 \ldots \frac{L}{2}$ and $j = \frac{L}{2} + 1 \ldots L$, and λ_{Id}). Then for each repeat we can assign an internal energy $\sum_{i=1}^{L/2} h_i(a_i) + \sum_{i=1}^{L/2} \sum_{j>i}^{L/2} J_{ij}(a_i, b_j)$ and a interaction energy $\sum_{i=1}^{L/2} \sum_{j=L/2+1}^{L} J_{ij}(a_i, b_j) + \lambda_{Id}$, which of course depends on the amino acids constituting each repeat.

On Fig 5A, we show the comparison between ΔΔG and the evolutionary energy calculated using Eq 4, done for three different ANK proteins: IκBα [45, 46], Notch [47] and p16 [48]. It should be noted that different experimental techniques return different values for ΔG for the same protein, non overlapping within experimental error, pointing that other factors contribute to the experimental quantification of ΔΔG. A linear fit returns $R^2 \approx 0.61$. From 152 mutations we analyzed, 124 (82%) are predicted favorable when the mutation stabilized the folding of the structure, and unfavorable when they have also been measured to destabilize. The predictions that deviated the most are mutations in Notch from Serine to Proline, which is a structural disruptor, and were not considered in the linear fit. A comparison against FoldX [49] predictions can be found on S6 Fig.

On Fig 5B, we show reported mutations on pp32 [50], a protein belonging to LRR family. Again, measurements with different methods report different values of ΔΔG. The linear fit returns a poor $R^2 \approx 0.21$, but 30 (75%) mutations are both predicted and reported unstabilizing.

A similar comparison was performed by [31] for small globular proteins with an expression related to Eq 1. To reduce the number of interaction parameters $J_{ij}(a_i, b_j)$ they explicitly used

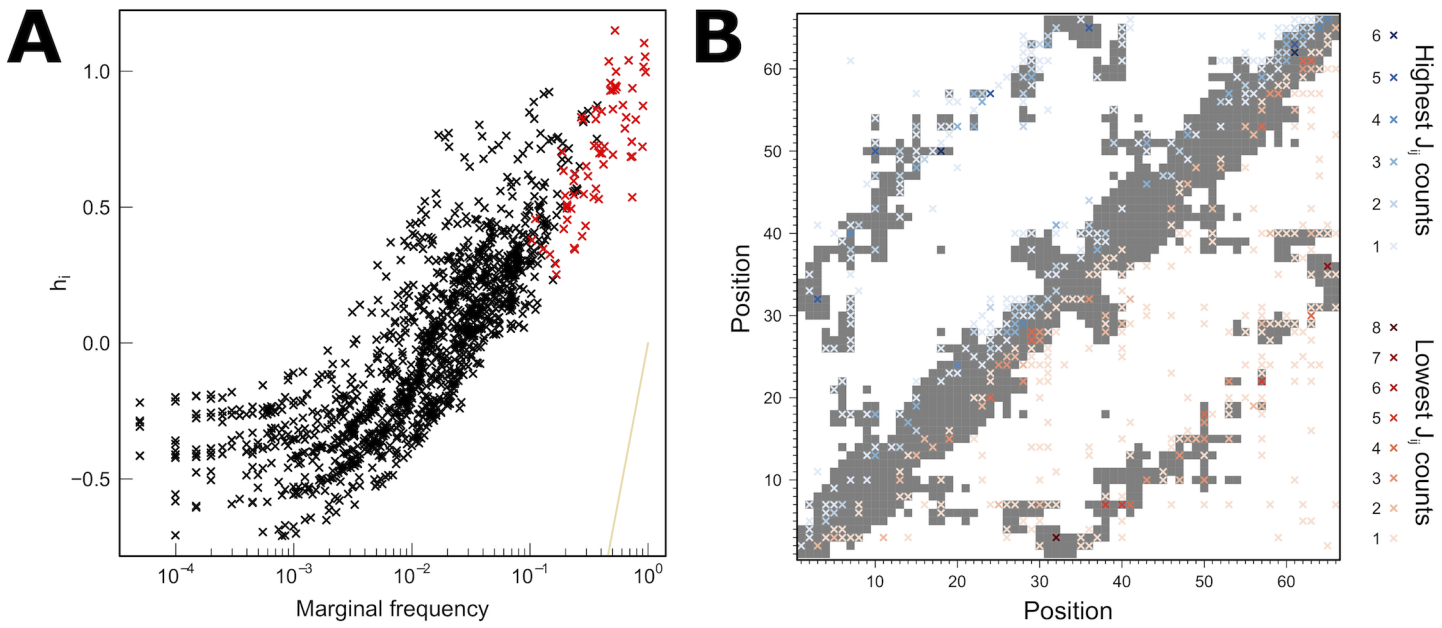

Fig 6. For the ANK family, on panel A we compare the parameters $h_i(a_i)$ to the marginal frequencies. The site-independent model (and initial condition) states that $h_i(a_i) = \ln(f_i(a_i))$. For the final model, this relation is tuned by the higher order correlations. On red, the parameters associated with the most common amino acid in each position are highlighted. On panel B, we compare the contact map of a pair of repeats of 1N0R (gray shadow) and the highest (blue) and lowest (red) $J_{ij}(a_i, b_j)$ parameters. The color scale indicates how many parameters involves the two positions (due to different sets of amino acids). Most extreme values fall into residues in contact or in the equivalent position of a repeat.

https://doi.org/10.1371/journal.pcbi.1005584.g006

structural information and set to zero all interactions between positions which are not in contact in the native structure. In contrast, we use a L_1-regularization to fix to zero those parameters which do not contribute significantly to the fitting process and obtain $J_{ij}(a_i, b_j) = 0$ and $J_{ij}(a_i, b_j) \neq 0$ in all pairs of positions, regardless they are supposed to be in contact or not in the 3D structure.

Interaction parameters are related to the structure and the sequence symmetry

Are the obtained parameters related to structural properties of these proteins? Local fields, $h_i(a_i)$, should account for the local propensity of each amino acid in each position, and therefore are expected to be related to $f_i(a_i)$. Fig 6A shows that the inferred $h_i(a_i)$ parameters are different from the initial condition $\ln(f_i(a_i))$ for the ANK family; that is, the values obtained for the parameters that account for higher order correlations are relevant. In red we highlight the points related to the consensus amino acid in each position. All of these residues have a strong local field associated to them, justifying why the construction of sequences with these amino acids results in foldable proteins. We also show a contact map of two ANK repeats (PDB id: 1N0R) on Fig 6B: gray background indicates that the two positions given by x and y axis are in contact in the native structure, and white that they are not. On the upper triangle of the figure and in blue crosses, we mark the positions involved in the highest J_{ij} parameters, i.e. those which imply higher coupling. A darker blue indicates that there are more J_{ij} (more combinations of amino acids) between those positions. Most of the highest J_{ij} match a pair of positions in contact in the 3D structure, or two which correspond to the same residue in the adjacent repeat patterns, i.e. i-th position in the first repeat and position j = i+33 in the second repeat. In red crosses we show the lowest J_{ij}, that mark a negative constraint. Again, a darker red

means that there are more J_{ij} with low values between those positions. It is apparent that these also involve mostly residues in contact, but shows that other regions are responsible for negative design.

Discussion

We propose a statistical model to account for fine details of the energy distribution in families of repeat proteins using only the sequences of amino acids. The model consists of a specialization of a Potts model to account for the local and pair-wise interactions and an extra term that includes higher order correlations, accounting for the similarity between consecutive repeats. The model is constrained by evolutionary characteristics of the families of proteins: we measure the frequencies of amino acids, co-occurrence of amino acids and the identity between repeats in extant natural proteins. To statistically define these quantities it is necessary to have a large set of sequences, which we showed are currently available for several repeat-protein families [37]. No information about the native folded conformation is required. The computation of the evolutionary energy field is computationally demanding, mostly due to long times spent in rigorous Monte Carlo simulations, but once the fitting is done the parameters can be used to score individual sequences fast and easily.

We studied three popular repeat protein families: ANK, TPR and LRR. After pre-processing of the alignments, we had enough sequences ($\approx$ 20500, 10000 and 18800 respectively) to fit the model to pairs of repeats of each family. We scored the *evolutionary energy* of all natural sequences in PFAM, and it allowed us to clearly distinguish between natural proteins and random sequences of amino acids: the first have energy values $<$ -50 and show a large spread while all random sequences have energy values ≈ 0. We evaluated designed repeat proteins which have been shown to fold and found that they score within the natural sequences distribution of energies. For the ANK and TPR families, these designed proteins have been shown to be highly stable upon thermal and chemical denaturation and, coincidentally, they are located at the most favorable side of the energy distribution of natural proteins, suggesting that the evolutionary energy score can be related to folding stability.

The energetic model can be used in Monte Carlo simulations to generate sequences that agree with the natural constraints of a given protein family. This ensemble of simulated sequences matches the amino acid frequencies, the identity between repeats and also the energy distribution of natural proteins. We found this set of simulated sequences is statistically indistinguishable from natural counterparts. Thus, the proposed model can be used as a tool to design repeat-protein sequences that have all the natural characteristics evaluated to date. Repeat proteins bind to other polypeptides and are candidates for specific binder scaffolds. Designed repeat proteins have been successfully synthesized and adapted to biomedical applications. Nevertheless, consensus design limits the possible variants as only a small proportion of residues are free to vary. Furthermore, they are extremely stable. Including coupling information can wide the possible sequences that can be studied, and could lead to more malleability of the designed molecules. Moreover, the stability change upon single point mutation can be well predicted by the model using just sequence information. For the ANK family, evolutionary energy variations correlate with the experimental values with an $R^2 \approx 0.6$. This improves FoldX [49] performance, which additionally requires a reference structure. Moreover, from the 152 experiments analyzed, the 82% predicts the direction of the stability change upon a point mutation. For the LRR family, the correlation is considerably lower, but 75% of the mutations are both predicted and reported in the available bibliography as destabilizing. For both the simulated sequences and for natural counterparts, we found that the similarity between consecutive repeats correlates with lower energy values, and that these are not

necessarily similar to the consensus sequence of the family, pointing out that duplication of stretches of sequences may well be an important factor in the evolution of these systems [51].

The existence of a simple and reliable energy function to score the 'evolutionary energy' of repeat-proteins can be used to trace the biological forces that acted upon their history, and to explore to which extent these conflict with the physical necessities of the systems [52]. Mapping the energy inhomogeneities along the repeat-arrays may allow us to infer the population of excited states in these proteins, many of which have been related to their physiological mechanisms.

Methods

Sequence alignments

Multiple sequence alignments of repeats were obtained from PFAM 27.0 [53]. The aligned sequences usually have misdetected initial and final residues. The amino acids at the ends of the repeat-detection do occur in the polypeptide chains (they are not actual deletions) and incorporating them improves the statistics of the real sequences. We completed these positions with the amino acids present on the actual proteins using the provided headers on the alignment and crossing information with UniProt database [54]. This leads to a reduction on the number of gaps in our alignments, which usually derives into noisy predictions in correlation analyses [31]. After, we created the alignment of pairs of repeats, joining sequences of repeats which are consecutive in a natural protein. Finally, we removed insertions from the alignments by deleting positions which have gaps in more than 80% of the sequences in the alignment.

Frequency calculations

Our model fits the occurrence of amino acids in every position, which we call the marginal frequency of residue a_i at position i of the alignment and denote $f_i(a_i)$, and the joint occurrence of two amino acids a_i and b_j simultaneously at two different positions of the alignment, $f_{ij}(a_i, b_j)$. To avoid biases by the overrepresentation of some proteins in the database, we used CD-HIT [55] to cluster sequences at 90% of identity and chose a representative sequence from each cluster. Finally, we computed by counting the $f_i(a_i)$ and $f_{ij}(a_i, b_j)$, and divided by the total number of sequences.

p_{id} calculations

From the same alignment explained in *Frequencies calculations*, for a sequence which has L residues constituting two consecutive repeats, the %Id between the repeats is the number of amino acids in positions i and $i + \frac{L}{2}$, for $i = 1 \ldots \frac{L}{2}$ which are exactly the same. Gaps are treated as an amino acid. Once we have the values for all sequences in an alignment, we define p_{id} as the proportion of sequences within the alignment with the same $\%Id$ between repeats.

Construction of an ensemble of sequences in agreement with a energy equation

Given a set of parameters h_i, J_{ij}, λ_{Id} and Eq 4, we use a Monte Carlo procedure and the Metropolis criterion to generate an ensemble of N sequences of length L each. We initiate with a random string of L residues. At each step, we produce a point mutation in any position. If this mutation is favorable, i.e. the energy is lower than that of the original sequence, we accept the mutation. If not, we accept the mutation with a probability of $e^{-\Delta E}$, where ΔE is the difference of energy between the original and the mutated sequence. When accepted, the mutated

sequence is used as the original one for next step. We add one sequence to our final ensemble every t steps (we used t = 1000).

Learning the parameters for the model

Our model is proposed to reproduce $f_i(a_i)$, $f_{ij}(a_i, b_j)$ and p_{id} from the alignment of natural sequences. To learn the set of parameters h_i, J_{ij}, λ_{Id} which reproduce them, we used a gradient descent procedure. In each step, an ensemble of N = 80000 sample sequences was produced via Monte Carlo using as energy the expression 4 and the trial parameters. We measured its marginal, joint frequencies and p_{id} and updated the local parameters according to:

$$h_i^{t+1} \leftarrow h_i^t - \epsilon_s \left[f_i(a_i) - f_i^{model}(a_i) \right] \tag{5}$$

As the number of parameters for coupling is large (= $21^2 L^2$), we used a regularization L_1 to force to 0 those parameters which are not contributing significantly to the modeled frequencies. Then, we update these parameters by:

$$J_{ij}^{t+1} \leftarrow \begin{cases} 0 & \text{if } J_{ij}^t = 0 \text{ and } |f_i(a_i,b_j) - f_i^{model}(a_i,b_j)| < \gamma \\ \epsilon_j \left[f_{ij}(a_i,b_j) - f_{ij}^{model}(a_i,b_j) - \gamma sign(f_{ij}(a_i,b_j) - f_{ij}^{model}(a_i,b_j)) \right] & \text{if } J_{ij}^t = 0 \text{ and } |f_{ij}(a_i,b_j) - f_{ij}^{model}(a_i,b_j)| > \gamma \\ J_{ij}^t + \epsilon_j \left[f_{ij}(a_i,b_j) - f_{ij}^{model}(a_i,b_j) - \gamma sign(J_{ij}^t) \right] & \text{if } \left[J_{ij}^t + \epsilon_j (f_{ij}(a_i,b_j) - f_{ij}^{model}(a_i,b_j) - \gamma sign(J_{ij}^t) \right] \cdot J_{ij}^t > 0 \\ 0 & \text{if } \left[J_{ij}^t + \epsilon_j (f_{ij}(a_i,b_j) - f_{ij}^{model}(a_i,b_j) - \gamma sign(J_{ij}^t) \right] \cdot J_{ij}^t < 0 \end{cases} \tag{6}$$

Finally, the parameters λ_{Id} are updated according to:

$$\lambda_{Id}^{t+1} \leftarrow \lambda_{Id}^t + \epsilon_{ID} \left[p_{id}(\%Id) - p_{id}^{model}(\%Id) \right] \tag{7}$$

We iterated until the maximum difference between the predicted frequencies and the natural sequences was below 0.02. This value was chosen according to the robustness of the frequencies estimations on the available data. We calculated the frequencies on half of the available sequences and compared the results to the frequencies counts on all the available sequences. The largest differences were slightly below 0.02. We believe that this maximum error thus reflects the actual error in the data and it is not reasonable to ask the model for more accuracy than that of the data itself. The code was written in C++ and is available at GitHub: https://github.com/proteinphysiologylab/2017_Espadaetal.

Supporting information

S1 Fig. Distribution of %Id of sequences generated by model on eq 1 (main text) in dotted lines and by model 4 (main text) in solid lines. In black dots, the natural sequences' distribution of %Id.
(PDF)

S2 Fig. Histogram of frequency errors.
(PDF)

S3 Fig. Change of J_{ij} parameters under regularization.
(PNG)

S4 Fig. Dendrogram based on pair-wise similarity. Natural and simulated sequences are indistinguishable from pairwise similarity.
(PDF)

S5 Fig. A) Logos for the MSA of natural pairs of ANK repeats (top), of simulated pairs of ANK repeats (center) and low energy pairs of simulated ANK repeats (bottom). B) Distribution of familiarity, as defined in Turjanski et al (2016) for the same sets of sequences. Simulated sequences reproduce the distribution of natural proteins and are indistinguishable.
(PNG)

S6 Fig. For the ANK family, we compared the results of our models on folding stability with FoldX, a popular tool used to predict stability changes upon mutation. As FoldX requires a reference structure, some of the constructions tested in the protein Notch cannot be analyzed, so we excluded them from our predictions for a fair comparison. It can be seen that, overall, our model (right panel) is a better predictor than foldX (left panel) of stability changes upon mutation, and has the advantage of using just the sequence information.
(PNG)

S7 Fig. At left, comparison between the local field parameters $h_i(a_i)$ and the marginal frequencies $f_i(a_i)$. At center, contact map (grey indicates position in contact, white not in contact on the native structure). On blue, pairs of positions involved in highest $J_{ij}(a_i, b_j)$, red lowest $J_{ij}(a_i, b_j)$. At right, pairs of amino acids involved in the highest $J_{ij}(a_i, b_j)$ parameters (on blue) and in the lowest $J_{ij}(a_i, b_j)$ (on red). ANK at the top. TPR at the center. LRR at the bottom.
(PDF)

Acknowledgments

R.G.P. is a long-term EMBO postdoctoral fellow (ALTF 212-2016). R.E. and R.G.P. would like to thank C.F.K. and D.U.F mentions MM.Cat

Author Contributions

Conceptualization: TM AMW DUF.

Data curation: RE RGP.

Formal analysis: RE RGP TM AMW DUF.

Funding acquisition: RE RGP TM AMW DUF.

Investigation: RE RGP TM AMW DUF.

Methodology: RE TM AMW DUF.

Project administration: RE RGP TM AMW DUF.

Resources: RE RGP TM AMW DUF.

Software: RE RGP.

Supervision: TM AMW DUF.

Validation: RE RGP TM AMW DUF.

Visualization: RE RGP TM AMW DUF.

Writing – original draft: RE RGP TM AMW DUF.

Writing – review & editing: RE RGP TM AMW DUF.

References

1. Main ER, Lowe AR, Mochrie SG, Jackson SE, Regan L. A recurring theme in protein engineering: the design, stability and folding of repeat proteins. Current opinion in structural biology. 2005; 15(4):464–471. https://doi.org/10.1016/j.sbi.2005.07.003 PMID: 16043339
2. Sedgwick SG, Smerdon SJ. The ankyrin repeat: a diversity of interactions on a common structural framework. Trends in biochemical sciences. 1999; 24(8):311–316. https://doi.org/10.1016/S0968-0004(99)01426-7 PMID: 10431175
3. Kobe B, Kajava AV. The leucine-rich repeat as a protein recognition motif. Current opinion in structural biology. 2001; 11(6):725–732. https://doi.org/10.1016/S0959-440X(01)00266-4 PMID: 11751054
4. Blatch GL, Lässle M. The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. Bioessays. 1999; 21(11):932–939. https://doi.org/10.1002/(SICI)1521-1878(199911)21:11%3C932::AID-BIES5%3E3.0.CO;2-N PMID: 10517866
5. Rowling PJ, Sivertsson EM, Perez-Riba A, Main ER, Itzhaki LS. Dissecting and reprogramming the folding and assembly of tandem-repeat proteins. Biochemical Society Transactions. 2015; 43(5):881–888. https://doi.org/10.1042/BST20150099 PMID: 26517898
6. Brunette T, Parmeggiani F, Huang PS, Bhabha G, Ekiert DC, Tsutakawa SE, et al. Exploring the repeat protein universe through computational protein design. Nature. 2015; https://doi.org/10.1038/nature16162 PMID: 26675729
7. Urvoas A, Guellouz A, Valerio-Lepiniec M, Graille M, Durand D, Desravines DC, et al. Design, production and molecular structure of a new family of artificial alpha-helicoidal repeat proteins (αRep) based on thermostable HEAT-like repeats. Journal of molecular biology. 2010; 404(2):307–327. https://doi.org/10.1016/j.jmb.2010.09.048 PMID: 20887736
8. Espada R, Sánchez IE, Ferreiro DU. Detailing Protein Landscapes under Pressure. Biophysical Journal. 2016; 111(11):2339–2341. https://doi.org/10.1016/j.bpj.2016.10.038 PMID: 27926834
9. Frauenfelder H. Function and Dynamics of Myoglobin. Annals of the New York Academy of Sciences. 1987; 504:151–167. https://doi.org/10.1111/j.1749-6632.1987.tb48730.x PMID: 3115167
10. Ferreiro DU, Komives EA, Wolynes PG. Frustration in biomolecules. Quarterly reviews of biophysics. 2014; 47(04):285–363. https://doi.org/10.1017/S0033583514000092 PMID: 25225856
11. Main ER, Xiong Y, Cocco MJ, D'Andrea L, Regan L. Design of stable α-helical arrays from an idealized TPR motif. Structure. 2003; 11(5):497–508. https://doi.org/10.1016/S0969-2126(03)00076-5 PMID: 12737816
12. Kajander T, Cortajarena AL, Mochrie S, Regan L. Structure and stability of designed TPR protein superhelices: unusual crystal packing and implications for natural TPR proteins. Acta Crystallographica Section D: Biological Crystallography. 2007; 63(7):800–811. https://doi.org/10.1107/S0907444907024353 PMID: 17582171
13. 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
14. Stumpp MT, Forrer P, Binz HK, Plückthun A. Designing repeat proteins: modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family. Journal of molecular biology. 2003; 332(2):471–487. https://doi.org/10.1016/S0022-2836(03)00897-0 PMID: 12948496
15. 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 computational biology. 2008; 4(5):e1000070. https://doi.org/10.1371/journal.pcbi.1000070 PMID: 18483553
16. Parra RG, Espada R, Sánchez IE, Sippl MJ, Ferreiro DU. Detecting repetitions and periodicities in proteins by tiling the structural space. The Journal of Physical Chemistry B. 2013; 117(42):12887–12897. https://doi.org/10.1021/jp402105j PMID: 23758291
17. Espada R, Parra RG, Sippl MJ, Mora T, Walczak AM, Ferreiro DU. Repeat Proteins challenge the concept of structural domains. Biochemical Society Transactions. 2015; 43(5):844–849. https://doi.org/10.1042/BST20150083 PMID: 26517892
18. Frauenfelder H, McMahon BH, Fenimore PW. Myoglobin: the hydrogen atom of biology and a paradigm of complexity. Proceedings of the National Academy of Sciences. 2003; 100(15):8615–8617. https://doi.org/10.1073/pnas.1633688100 PMID: 12861080
19. 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 PMID: 22106262

20. 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 PMID: 23410359
21. Ekeberg M, Hartonen T, Aurell E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. Journal of Computational Physics. 2014; 276:341–356. https://doi.org/10.1016/j.jcp.2014.07.024
22. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ. Learning generative models for protein fold families. Proteins: Structure, Function, and Bioinformatics. 2011; 79(4):1061–1078. https://doi.org/10.1002/prot.22934 PMID: 21268112
23. Morcos F, Jana B, Hwa T, Onuchic JN. Coevolutionary signals across protein lineages help capture multiple protein conformations. Proceedings of the National Academy of Sciences. 2013; 110 (51):20533–20538. https://doi.org/10.1073/pnas.1315625110 PMID: 24297889
24. Cheng RR, Morcos F, Levine H, Onuchic JN. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. Proceedings of the National Academy of Sciences. 2014; 111(5):E563–E571. https://doi.org/10.1073/pnas.1323734111 PMID: 24449878
25. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure prediction. Proceedings of the National Academy of Sciences. 2012; 109(26):10340–10345. https://doi.org/10.1073/pnas.1207864109 PMID: 22691493
26. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. Elife. 2014; 3:e02030. https://doi.org/10.7554/eLife.02030 PMID: 24842992
27. Cheng RR, Raghunathan M, Noel JK, Onuchic JN. Constructing sequence-dependent protein models using coevolutionary information. Protein Science. 2016; 25(1):111–122. https://doi.org/10.1002/pro.2758 PMID: 26223372
28. Zahnd C, Wyler E, Schwenk JM, Steiner D, Lawrence MC, McKern NM, et al. A designed ankyrin repeat protein evolved to picomolar affinity to Her2. Journal of molecular biology. 2007; 369(4):1015–1028. https://doi.org/10.1016/j.jmb.2007.03.028 PMID: 17466328
29. Cliff MJ, Williams MA, Brooke-Smith J, Barford D, Ladbury JE. Molecular recognition via coupled folding and binding in a TPR domain. Journal of molecular biology. 2005; 346(3):717–732. https://doi.org/10.1016/j.jmb.2004.12.017 PMID: 15713458
30. Levy RM, Haldane A, Flynn WF. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. Current Opinion in Structural Biology. 2017; 43:55–62. https://doi.org/10.1016/j.sbi.2016.11.004 PMID: 27870991
31. 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
32. Sutto L, Marsili S, Valencia A, Gervasio FL. From residue coevolution to protein conformational ensembles and functional dynamics. Proceedings of the National Academy of Sciences. 2015; 112 (44):13567–13572. https://doi.org/10.1073/pnas.1508584112 PMID: 26487681
33. Haldane A, Flynn WF, He P, Vijayan R, Levy RM. Structural Propensities of Kinase Family Proteins from a Potts Model of Residue Co-Variation. Protein Science. 2016; https://doi.org/10.1002/pro.2954 PMID: 27241634
34. 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
35. Hopf, TA, Ingraham, JB, Poelwijk, FJ, Springer, M, Sander, C, Marks, DS. Quantification of the effect of mutations using a global probability model of natural sequence variation. arXiv preprint arXiv:151004612. 2015;.
36. Schüler A, Bornberg-Bauer E. Evolution of Protein Domain Repeats in Metazoa. Molecular Biology and Evolution. 2016; https://doi.org/10.1093/molbev/msw194 PMID: 27671125
37. 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
38. 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 PMID: 12461176

39. 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 PMID: 19116270

40. Finkelstein AV, Badretdinov AY, Gutin AM. Why do protein architectures have boltzmann-like statistics? Proteins: Structure, Function, and Bioinformatics. 1995; 23(2):142–150. https://doi.org/10.1002/prot.340230204 PMID: 8592696

41. Krachler AM, Sharma A, Kleanthous C. Self-association of TPR domains: Lessons learned from a designed, consensus-based TPR oligomer. Proteins: Structure, Function, and Bioinformatics. 2010; 78 (9):2131–2143. https://doi.org/10.1002/prot.22726 PMID: 20455268

42. Cortajarena AL, Wang J, Regan L. Crystal structure of a designed tetratricopeptide repeat module in complex with its peptide ligand. FEBS journal. 2010; 277(4):1058–1066. https://doi.org/10.1111/j.1742-4658.2009.07549.x PMID: 20089039

43. Zhu H, Sepulveda E, Hartmann MD, Kogenaru M, Ursinus A, Sulz E, et al. Origin of a folded repeat protein from an intrinsically disordered ancestor. eLife. 2016; 5:e16761. https://doi.org/10.7554/eLife.16761 PMID: 27623012

44. Turjanski P, Parra RG, Espada R, Becher V, Ferreiro DU. Protein Repeats from First Principles. Scientific reports. 2016; 6. https://doi.org/10.1038/srep23959 PMID: 27044676

45. DeVries I, Ferreiro DU, Sánchez IE, Komives EA. Folding kinetics of the cooperatively folded subdomain of the IκBα ankyrin repeat domain. Journal of molecular biology. 2011; 408(1):163–176. https://doi.org/10.1016/j.jmb.2011.02.021 PMID: 21329696

46. Ferreiro DU, Cervantes CF, Truhlar SM, Cho SS, Wolynes PG, Komives EA. Stabilizing IκBα by "consensus" design. Journal of molecular biology. 2007; 365(4):1201–1216. https://doi.org/10.1016/j.jmb.2006.11.044 PMID: 17174335

47. Street TO, Bradley CM, Barrick D. An improved experimental system for determining small folding entropy changes resulting from proline to alanine substitutions. Protein science. 2005; 14(9):2429–2435. https://doi.org/10.1110/ps.051505705 PMID: 16131666

48. Tang KS, Fersht AR, Itzhaki LS. Sequential unfolding of ankyrin repeats in tumor suppressor p16. Structure. 2003; 11(1):67–73. https://doi.org/10.1016/S0969-2126(02)00929-2 PMID: 12517341

49. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic acids research. 2005; 33(suppl 2):W382–W388. https://doi.org/10.1093/nar/gki387 PMID: 15980494

50. Dao TP, Majumdar A, Barrick D. Highly polarized C-terminal transition state of the leucine-rich repeat domain of PP32 is governed by local stability. Proceedings of the National Academy of Sciences. 2015; 112(18):E2298–E2306. https://doi.org/10.1073/pnas.1412165112 PMID: 25902505

51. Björklund ÅK, Ekman D, Elofsson A. Expansion of protein domain repeats. PLoS Comput Biol. 2006; 2 (8):e114. https://doi.org/10.1371/journal.pcbi.0020114 PMID: 16933986

52. Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. Proceedings of the National Academy of Sciences. 2014; 111(34):12408–12413. https://doi.org/10.1073/pnas.1413575111 PMID: 25114242

53. 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

54. Consortium U, et al. UniProt: the universal protein knowledgebase. Nucleic acids research. 2017; 45 (D1):D158–D169. https://doi.org/10.1093/nar/gkw1099 PMID: 27899622

55. 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