PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Evolution and folding of repeat proteins

Ezequiel A. Galpern[a,b], Jacopo Marchi[c,1,2], Thierry Mora[c], Aleksandra M. Walczak[c], and Diego U. Ferreiro[a,b,3]

Edited by Vincenzo Carnevale, Temple University, Philadelphia, PA; received March 8, 2022; accepted June 22, 2022 by Editorial Board Member William F. DeGrado

Repeat proteins are made with tandem copies of similar amino acid stretches that fold into elongated architectures. These proteins constitute excellent model systems to investigate how evolution relates to structure, folding, and function. Here, we propose a scheme to map evolutionary information at the sequence level to a coarse-grained model for repeat-protein folding and use it to investigate the folding of thousands of repeat proteins. We model the energetics by a combination of an inverse Potts-model scheme with an explicit mechanistic model of duplications and deletions of repeats to calculate the evolutionary parameters of the system at the single-residue level. These parameters are used to inform an Ising-like model that allows for the generation of folding curves, apparent domain emergence, and occupation of intermediate states that are highly compatible with experimental data in specific case studies. We analyzed the folding of thousands of natural Ankyrin repeat proteins and found that a multiplicity of folding mechanisms are possible. Fully cooperative all-or-none transitions are obtained for arrays with enough sequence-similar elements and strong interactions between them, while noncooperative element-by-element intermittent folding arose if the elements are dissimilar and the interactions between them are energetically weak. Additionally, we characterized nucleation-propagation and multidomain folding mechanisms. We show that the global stability and cooperativity of the repeating arrays can be predicted from simple sequence scores.

repeat proteins | protein folding | co-evolution | Ising

Significance

Protein sequences change over evolutionary times. The fixation of these changes is coupled to the protein's capability to properly explore conformational space and perform biological functions. Repeat proteins are privileged systems to understand the relations between evolution and folding, due to their conserved structure and functional diversity. Here, we use the evolutionary record of natural sequences to model and analyze the folding mechanisms of thousands of proteins. The proposed model successfully reproduces folding experiments using only sequence information as input. We performed large-scale predictions of folding mechanisms in the most abundant repeat-protein family, identifying higher-order features such as domain emergence, stability, and cooperativity of repeat arrays.

Robust folding and long-term evolution are two of the most basic aspects of natural proteins. These features are necessarily intertwined, as the sequences we find today are the result of selection of specific instances that, when folded, minimize the energetic conflicts between their amino acids: They are overall "minimally frustrated" heteropolymers (1). The energy-landscape theory of protein folding recognizes these fundamental aspects and shows that the general topography of the energy landscape of globular domains is that of a rough funnel, in which the native interactions are, on average, more favorable than nonnative ones. In accordance, the population of the folding routes can be reasonably well predicted with topological models of the native state (2), and, for most globular domains, local energetic differences rarely perturb the global aspects of the folding mechanisms (3). This is not the typical situation in the case of repeat proteins.

Repeat proteins are composed of tandem arrays of similar amino acid stretches. The repeats usually fold in recursive structural elements that pack against each other in a roughly periodic way, making the overall architecture of the arrays appear as elongated objects (4). In these, folding domains are not easy to define and identify, as several, but not necessarily all, of the repetitions cooperate in the stabilization of structures (5). Being quasi-one-dimensional, the folding of the complete array is dominated by the local energetics within each repeat and its local neighbors, making the folding sensitive to small perturbations that may lead to the breakdown of cooperativity and the appearance of stable intermediates and subdomains (5). Notably, simple coarsed one-dimensional Ising-like models of repeat protein have been found to be extremely useful for interpreting in vitro experiments (6). In general, the folding mechanisms are defined by an initial nucleation in some region of the array and the propagation of structure to their neighbors. When the local energetics are similar along the assemblage, parallel folding routes can be identified (7), and the routes can be switched by (de)stabilizing regions along the array (8). Thus, the energy landscape of repeat proteins appears "plastic" and very amenable to design (9). To what extent nature has exploited this opportunity is yet unknown.

Besides single-point mutations, the evolution of repeat proteins is thought to occur via duplications and deletions of large portions of primary structure, usually encompassing one or more repeats (10). These proteins are present in all taxa and are particularly abundant in eukaryotes, where they account for about 20% of the coded proteins. Their activity is usually associated with specific protein–protein interactions, with a versatility that can be equated to that of antibodies. In various cases, the detailed folding

Author contributions: E.A.G., T.M., A.M.W., and D.U.F. designed research; E.A.G., J.M., T.M., A.M.W., and D.U.F. performed research; E.A.G., J.M., T.M., and A.M.W. contributed new reagents/analytic tools; E.A.G., T.M., A.M.W., and D.U.F. analyzed data; E.A.G., T.M., A.M.W., and D.U.F. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. V.C. is a guest editor invited by the Editorial Board.

[1]Present address: School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332.

[2]Present address: Institut de Biologie, École Normale Supérieure, F-75230 Paris, France.

[3]To whom correspondence may be addressed. Email: ferreiro@qb.fcen.uba.ar.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2204131119/-/DCSupplemental.

Published July 29, 2022.

mechanism of the repeat arrays has been identified to play a major role in their biological function (11), but for most of the repeat arrays, it remains unknown. Here, we aim to use evolutionary information from repeat-protein systems to investigate the folding mechanisms of thousands of natural repeat arrays. We will make use of Ankyrin repeat proteins, as this is one of the most abundant families, and their folding mechanism can be well approximated with simple folding models (12). We hypothesize that the local energetics can be estimated with a maximum entropy model for the natural sequence statistics that results in a pairwise Potts model for amino acid interactions (13). We map the energetics of the sequences to an Ising model with one free parameter that we fit with experimental folding data. The resulting model is then applied to thousands of different sequences that fold to the same overall topology, revealing distinct folding routes, the emergence of subdomains, downhill scenarios, etc. in a variegated zoo of folding mechanisms. We found that the overall folding behavior of a complete repeat array can be well described with few global descriptors that can be directly calculated solely from sequence information.

Results

Model Definition. We consider a repeat protein as a tandem array of consecutive folding elements, each of which can be either folded (F) or unfolded (U). Each element corresponds to a group of consecutive amino acids, whose behaviors are considered together as one spin variable. The most simple assignment is to match a whole repeat with one spin, but the model can be generalized considering repeat subunits consistently. Specific interactions take place between neighbor elements if both are folded. Therefore, the system can be represented by a finite Ising chain with N elements, where the energy of a coarse-grained configuration, the Hamiltonian, is given by the free energy of the corresponding ensemble of microstates

$$H = -\sum_{j=1}^{N}\left[Ts_j\left(1-\delta_{j,F}\right)+\epsilon_j^i\delta_{j,F}\right]-\sum_{j=1}^{N-1}\sum_{k>j}\epsilon_{jk}^s\delta_{j,F}\delta_{k,F}, \qquad [1]$$

where $\delta_{j,F}$ is the Kronecker symbol, taking value one if element j is folded and zero otherwise. Hence, if j is unfolded, there is an explicit contribution to the free energy given by the entropy s_j of the available spatial configurations of the element, and we take the contributions to the internal energy to be zero. If the element is folded ($\delta_{j,F}=1$), we approximate the native state to be compact enough; hence, there is no intrinsic entropy contribution, but an internal energy ϵ_j^i is assigned. Two elements interact with a surface energy ϵ_{jk}^s only if both are folded. A similar coarse-grained repeat-protein Ising model has been exhaustively studied by using arbitrary parameters and compared to molecular dynamics simulations for the TPR family (14). Here, we consider that the internal and surface energy parameters are functions of the amino acid sequence. Focusing on Ankyrins, a single family in which a native repeated structure is roughly conserved (15), we hypothesize that sequence variation within repeat units is linked to changes in local stability. Hence, in order to calculate ϵ_j^i and ϵ_{jk}^s for a sequence, we used residue–residue couplings and local fields that have been inferred for the Ankyrin family using a combination of a Direct Coupling Analysis (DCA) (16, 17) and an explicit mechanistic evolution scheme of whole-repeats duplications and deletions (details are in *SI Appendix*). Given a sequence, the evolutionary statistical function (often called energy) is given by a Potts model. For example, for a sequence $\boldsymbol{\sigma}$ with two repeats of L residues, Potts energy can be written as

$$E(\boldsymbol{\sigma}) = -\sum_{a=1}^{2L}\tilde{h}_a(\sigma_a)-\sum_{a,b=1}^{2L}\tilde{J}_{ab}(\sigma_a,\sigma_b), \qquad [2]$$

and can be generalized to an arbitrarily long array of N repeats. Please note that a simplified notation is used here; a detailed notation is included in *SI Appendix*.

This kind of statistical energy has been reported to predict evolutionary features (18), as well as some fitness effect or global stability change given by point mutations (13, 19–21). Indeed, we compared the experimental folding-energy difference between mutants and wild type ($\Delta\Delta G$) available in the literature for natural proteins of the Ankyrin family (8, 22–27), and we compute ΔE for the same mutants of three ANK proteins, finding a linear trend (*SI Appendix*, Fig. S7, $R^2\simeq 0.6$). More details are provided in *SI Appendix*, *SI Methods*. If we assume that there is no entropy difference between point mutants, the coarse-grained folding energy of fragments can be calculated simply by locally applying evolutionary fields $\tilde{h}_a$ and $\tilde{J}_{ab}$ to a sequence $\boldsymbol{\sigma}$ and rescaling properly. We define ϵ_j^i and ϵ_{jk}^s as explicit functions of $\boldsymbol{\sigma}_j$ and $\boldsymbol{\sigma}_k$, the sequences of the folding elements j and k

$$\epsilon_j^i=\epsilon^i(\boldsymbol{\sigma}_j)=\frac{1}{\alpha}\left[\sum_{a\in j}\tilde{h}_a(\sigma_a)+\sum_{a,b\in j}\tilde{J}_{ab}(\sigma_a,\sigma_b)\right], \qquad [3a]$$

$$\epsilon_{jk}^s=\epsilon^s(\boldsymbol{\sigma}_j,\boldsymbol{\sigma}_k)=\frac{1}{\alpha}\left[\sum_{\substack{a\in j\\ b\in k}}\tilde{J}_{ab}(\sigma_a,\sigma_b)\right], \qquad [3b]$$

where $a\in j$ means the sequence position a is in the folding element j and $\alpha=-1.3$ is the fitted slope in *SI Appendix*, Fig. S7. The fields $\tilde{h}$, $\tilde{J}$ are set to give on average zero evolutionary energy E for random sequences—i.e., successions of amino acids (or gaps) sampled from a uniform distribution. Therefore, in this model, ϵ^i and ϵ^s do not contribute, on average, to the folding energy for random sequences and maximize for a configuration that minimizes E, as $\alpha<0$.

We present p16 protein as a example of the model definitions in Fig. 1*A*. We choose the folding elements to be sequence fragments of 18 and 15 sites, roughly coinciding with the sequence of each alpha helix in the typical Ankyrin repeat structure, as it was done previously (28). As in the evolution model (Eq. **2**), interactions were allowed between residues within a repeat and between first-neighbor repeats; in the folding model (Eq. **1**), each helix interacts with the other one in the same repeat and with the next two. Therefore, interactions farther than the nearest neighbor are allowed between folding elements. This is shown in Fig. 1*B*, where coevolutionary fields applied to a sequence are represented [$\tilde{J}_{ab}(\sigma_a,\sigma_b)$ as a matrix and $\tilde{h}_a(\sigma_a)$ on its diagonal], highlighting evolutionarily favorable or unfavorable residues and pairs of residues. In this case, as in general, the partial sums of these contributions give almost all favorable folding energy terms (Eq. **3**).

The scheme we propose leaves the intrinsic configurational entropy s_j undefined. We simplified its definition taking s_j to be independent of amino acid identity and strictly additive, therefore, for an element j with L_j residues $s_j=L_j\,s$. Hence, the single free parameter of the model is s, an average residue contribution to the effective number of configurations that a folding element polymer can access when unfolded.

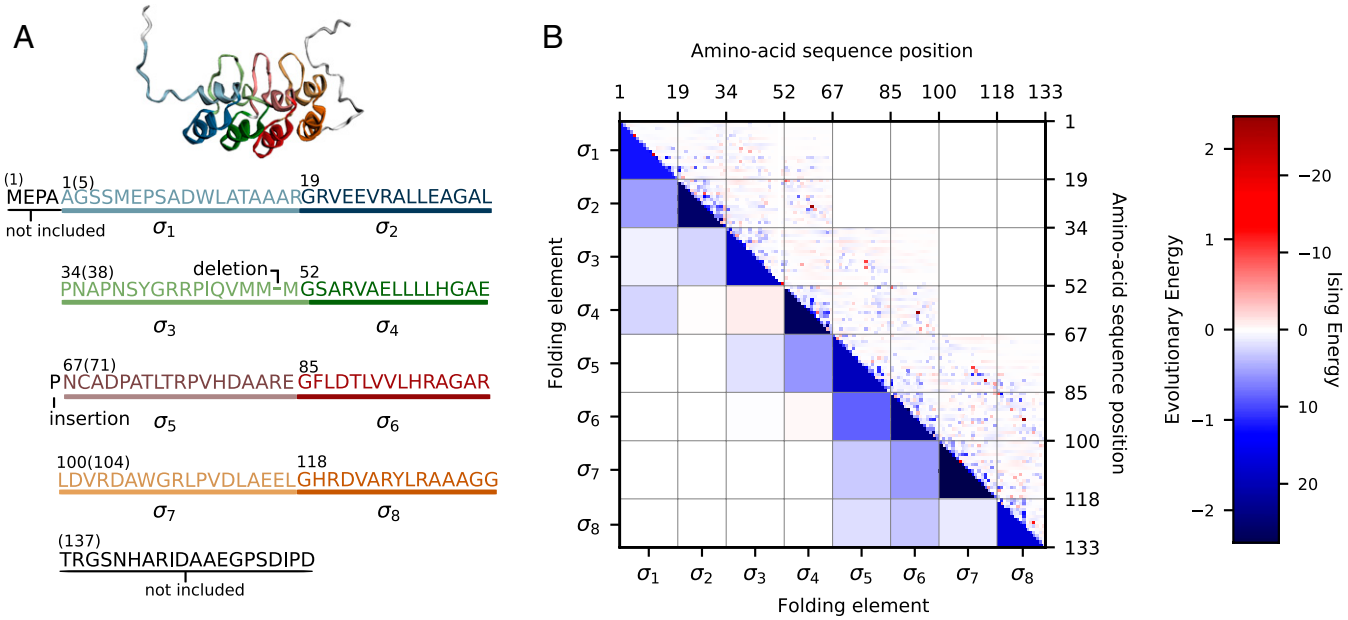

Fig. 1. Model definition. (*A*) Folding elements are 18- and 15-residue repeat subunits, as highlighted on the p16 sequence. They correspond to first and second alpha-helix of each repeat, respectively, as colored on the Protein Data Bank (PDB) structure. (*B*) Double energy heat map for p16. Evolutionary residue pairwise couplings $\tilde{J}_{ab}(\sigma_a, \sigma_b)$ on the upper side and single-residue contributions $\tilde{h}_a(\sigma_a)$ on the diagonal were added according to folding elements and rescaled to define Ising energies $\epsilon^i(\boldsymbol{\sigma}_j)$ and $\epsilon^s(\boldsymbol{\sigma}_j, \boldsymbol{\sigma}_k)$ on the lower side. Blue (red) dots represent evolutionary (un)favorable pairs and positions. On the lower side, almost all internal and interaction folding energies are favorable; hence, they are blue.

Case Studies. We ran Monte Carlo simulations for a well-studied four-repeat ANK protein, the CDK4/CDK6 inhibitor p16. The fraction of folded elements as a function of temperature was compatible with an experimental Circular Dichroism (CD) signal obtained from the literature (23) (Fig. 2*A*). For the same entropy per residue s, the effect of a point mutation in the unfolding curve was precisely reproduced (*SI Appendix*, Fig. S9). Nevertheless, at low temperature T (190 to 250 K), simulated curves also showed a less cooperative pretransition (Fig. 2*A*).

We define apparent folding domains as groups of elements, not necessarily consecutive in structure, that undergo folding transitions together. A mathematical definition is described in *Materials and Methods*. For p16, this is the case of elements 4 to 7, which actually behave collectively in an apparent folding domain. Fig. 2*C* and *D* shows that the first domain to fold (the nucleus) is closely followed by element 2 (a single-element domain). Interestingly, element 3, which behaves separately presenting low stability, corresponds to the first half of the second repeat, which

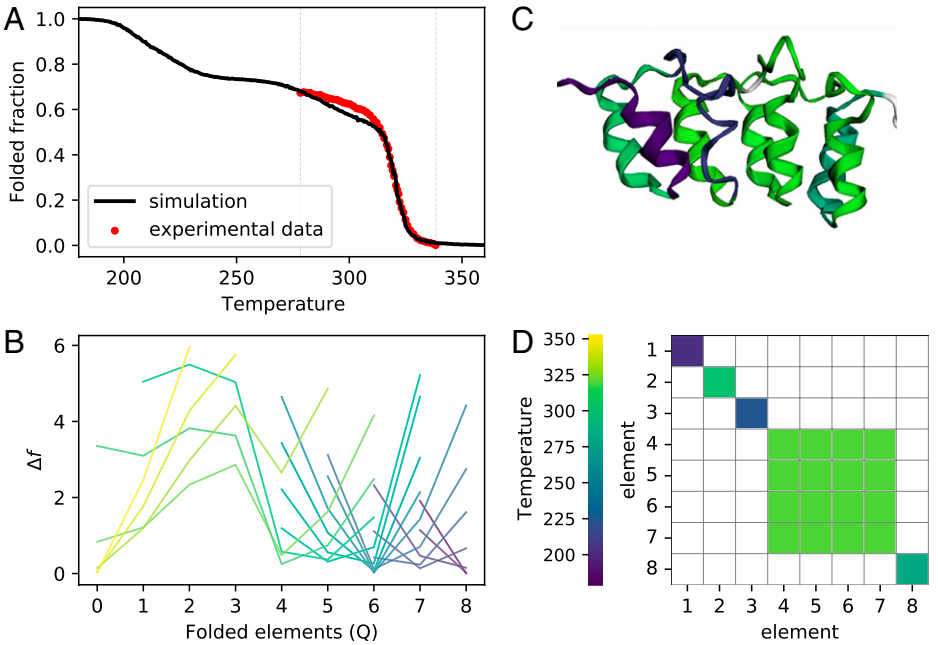

Fig. 2. Simulation results for p16. (*A*) Simulated (black) and experimental (red) thermal unfolding curves. Vertical gray dotted lines show the temperature range of the experiment. (*B*) Free-energy profiles, colored by temperature (same of *C* and *D*), with the number of folded elements Q as reaction coordinate. There is an all-or-none transition from $Q = 0$ to $Q = 4$ with a barrier in between; then, the minimum moves without any barrier. (*C*) The PDB structure is colored according to the folding temperature of each element. Purple fragments are the most unstable ones. (*D*) Apparent domain matrix, colored by domain folding temperatures. The first domain to fold (elements 4 to 7) corresponds to the all-or-none transition described in *B*, consistently with a nucleation-propagation mechanism.

has been found by NMR to fold as a turn instead of a helix (29, 30). Consistently, molecular dynamics simulations have shown in this region significant fluctuations in the folded state, and it is believed to be functionally relevant for binding (31). The first and last helices also presented low folding temperatures. In addition to the border effect, we remark that the evolutionary model was learned from internal repeats because terminal ones were considered different biological objects that can present modifications in sequence when compared to internal repeats (15, 32).

A free-energy profile $\Delta f(Q)$, where Q is the number of folded elements, highlights a nucleation-propagation mechanism (Fig. 2*B*). The nucleation of elements 4 to 7 is an all-or-none transition from $Q = 0$ to $Q = 4$ with a free-energy barrier in between. Structure then propagates visiting every remaining Q one by one as temperature is lowered. As a measure of cooperativity, we defined a score $\rho = Q_{barrier}/(N-1) = 3/7$, the fraction of intermediary Q that were not a minimum of $\Delta f(Q)$ for any T in a protein with N elements.

In addition to p16, we performed similar analyses on other natural ANK-containing proteins with available experimental folding data. We fit s to reproduce the reversible CD thermal unfolding curves of TRPV4 (33), TANC1 (34), and Kidney ANK 1 (35), finding optimal values in the range between 4.2 and 6.2 $\text{cal}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}\cdot\text{res}^{-1}$ (including p16; *SI Appendix*, Fig. S9). This interval is consistent with the calculations made by Baxa et al. (36) (4.3 $\text{cal}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}\cdot\text{res}^{-1}$ for an 18-residue fragment), it overlaps with the range estimated by D'Aquino et al. (37) (3.6 to 10.5 $\text{cal}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}\cdot\text{res}^{-1}$), and it is lower than the Makhatadze and Privalov (38) proposal ($\sim$11 $\text{cal}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}\cdot\text{res}^{-1}$). Given the range we obtained from experimental data fits, we set $s = 5$ $\text{cal}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}\cdot\text{res}^{-1}$ from here on, allowing us to perform simulations where thermal unfolding data are unavailable.

Consistent with reported folding dynamics (39–41), our analysis on IκBα (*SI Appendix*, Fig. S4) and *Drosophila melanogaster* Notch Receptor (*SI Appendix*, Fig. S5) allowed us to precisely identify highly cooperative domains where folding starts and distinguish them from less stable, folding-on-binding or flexible regions. On the other hand, the description of the AnkyrinR D34 24-element fragment unfolding via a stable intermediate with an unstructured half (42) was not reproduced (*SI Appendix*, Fig. S6). In this case, it is possible that considering the entropy strictly additive underestimates the probability of some partially folded states. In addition, our model does not allow for alternative configurations that may account for the local frustration that has been reported to correlate with the formation of the intermediate in D34 (43). A detailed analysis of these cases is provided in *SI Appendix*, *Case Studies*.

Although trained on natural sequences, the model can also be applied on Designed Ankyrin Repeat proteins (DARPins). Reported experimental thermal unfolding CD signal was compatible with simulation results in two four-repeat sequences, but three-repeat DARPins curves were not as close to the experimental ones for the same s (*SI Appendix*, Fig. S9 *E* and *F*). For a library of 100 DARPins generated by using Plückthun's framework (44), we found that stability increased with repeat-array length (*SI Appendix*, Fig. S10), consistent with previous reports (45). In summary, the statistical model learned on natural sequences reproduced globally the folding temperature of consensus-designed proteins, but the local surface and internal energies assignment is not sufficient to account for all the observed folding transitions of these artificial proteins.

General Results. To evaluate the folding of thousands of repeat proteins, we applied the model on a selected subset with 4,020 natural sequences formed with 4 to 36 repeats (8 to 72 elements) with no insertions or deletions. For each of these, we computed the thermal unfolding curves, the apparent domains, and the free-energy profiles, and we found a large variety of folding behaviors. The dataset included short proteins with strictly two-state transitions and a single domain as A2F665 (Fig. 3*A*, $\rho = 1$)

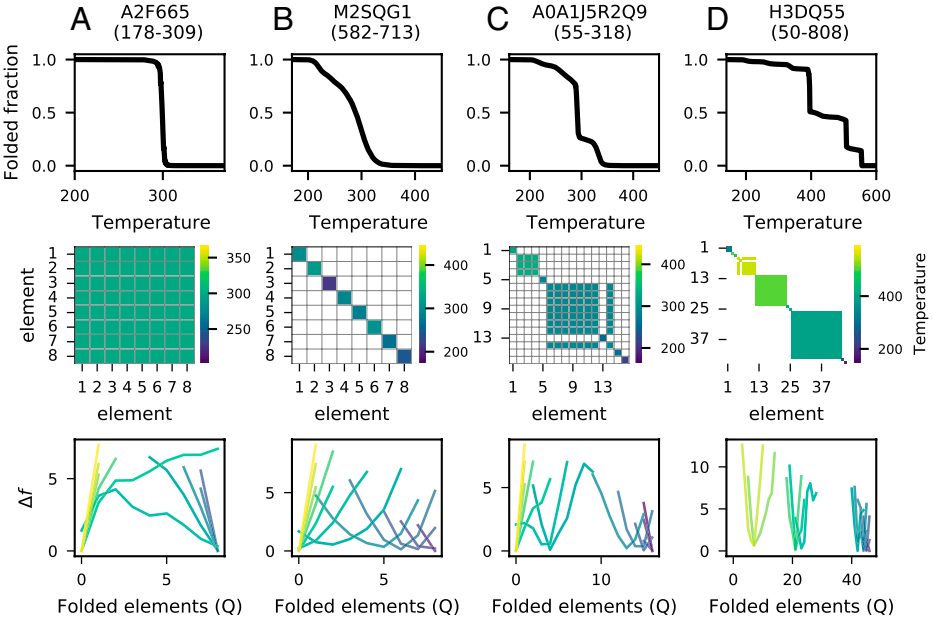

Fig. 3. Different folding mechanisms. Thermal unfolding curves (*Top*), apparent domain matrix and temperature scale (*Middle*), and free-energy profiles (*Bottom*) for four proteins are shown. We identified each sequence with the UniProt (46) code and the position range of the ANK repeat array. (*A*) A2F665 (178 to 309) had a highly cooperative two-state transition and a single domain ($\rho = 1$). (*B*) M2SQG1 (582 to 713) did not present any free-energy barrier, but elements unfolded one by one uncooperatively ($\rho = 0$). (*C*) Eight-repeat (16 folding elements) A0A1J5R2Q9 (55 to 318) presented many domains of different sizes and an unfolding curve with pretransitions and posttransitions ($\rho = 0.67$). (*D*) H3DQ55 (50 to 808) is a long protein of 46 folding elements that folds in three steps ($\rho = 0.84$).

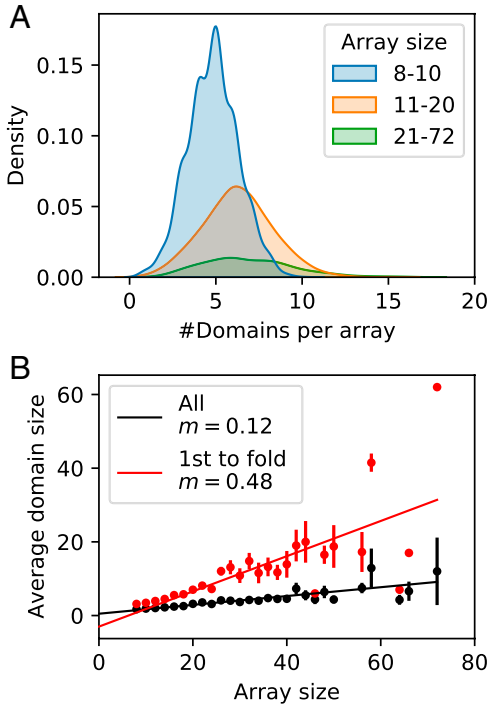

Fig. 4. Domain statistics. (*A*) Domain count per array histogram for short (blue), intermediate (orange), and long (green) arrays. (*B*) Average domain size as a function of array size for all (black) and only for the first domain to fold (red). Error bars are the SEs. Linear fit slopes *m* represented the average fraction of the array covered by a domain, which were 48% for the first to fold and 12% for all.

and others like M2SQG1 (Fig. 3*B*), which presented downhill folding without any free-energy barrier ($\rho = 0$). As protein size increased, we found multidomain examples (Fig. 3*C* and *D*), where several nucleations and propagations appeared. Large 46-element H3DQ55 (Fig. 3*D*) presented a step-like folding, with wide stability gaps between large domain-like all-or-none transitions. Terminal-repeats distinct behavior is widespread along the dataset, being more relevant in short proteins, where its effect can represent up to 50% of the fraction folded than in longer ones.

There is no characteristic size for the folding domains that emerges independently of protein size. On the contrary, the proteins spontaneously fold, on average, in 5.5 apparent folding domains, with a minor shift in the distribution when their full length is considered (Fig. 4*A*). Hence, domain size grows with protein size, and, on average, each domain covers 12% of the tandem array: While short proteins with 8 folding elements, on average, present less than 2-element domains, a long array with more than 40 folding elements typically presents 10-element domains. If we consider for each protein only the first domain to fold (the nucleus), the trend with array size becomes stronger. The nucleus domain represents, on average, 48% of the sequence (Fig. 4*B*). Also, short domains present a wide range of folding temperatures, while longer ones (with five or more elements) show more stability. The nucleus domains had also a roughly exponential dependence with size, but mostly fold at physiological or higher temperatures (*SI Appendix*, Fig. S11).

Protein folding temperature T_f and cooperativity score ρ did not correlate with each other (*SI Appendix*, Fig. S12), suggesting that the complexity of the system requires at least these two parameters to define the folding dynamics. Nevertheless, conditioning on array length revealed that there is a region of high cooperativity and stability where long arrays are more concentrated (*SI Appendix*, Fig. S13). Interestingly, it has been shown that these large repeat proteins are naturally formed with similar and energetically favorable elements (32).

Ankyrin domain proteins are reported to play a role in a variety of biological activities. We scanned all the Gene Ontology annotations with experimental evidence on the full dataset proteins and computed, for the corresponding sequences, T_f and ρ, but we did not find any clear relation between them and molecular function (*SI Appendix*, Fig. S14). Thus, the annotated function in the Gene Ontology is not a simple emergent of the folding properties of the sequences.

Model Interpretation. How exactly are stability and cooperativity related to sequence statistics? On one hand, protein T_f estimations were highly correlated with length-normalized coevolutionary energy of the respective sequences (Fig. 5*A*). This general agreement between folding stability and the global evolutionary statistical energy was an expected output of the mapping between the evolutionary and the folding model, given the definitions of Eq. **3**, and it is consistent with experimental results for other protein families (47).

On the other hand, cooperativity has a more complex dependence with coevolutionary energy. For a given protein, if differences between folding Ising internal energies ϵ_j^i of interacting

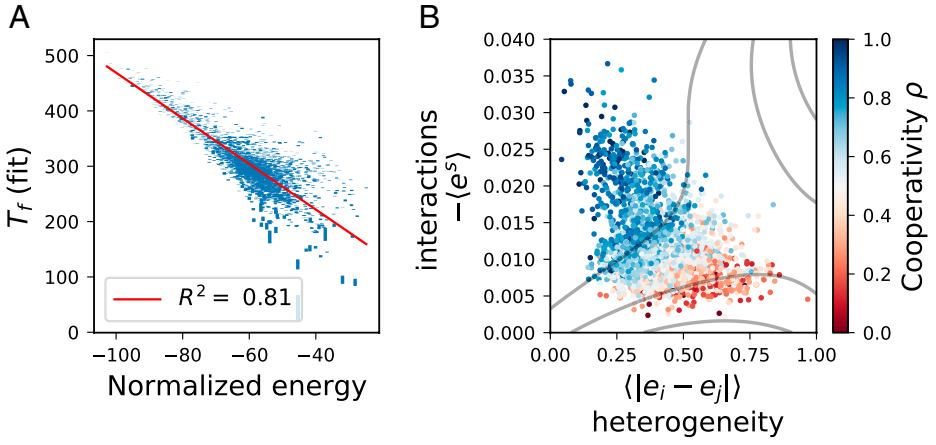

Fig. 5. Model interpretation. (*A*) Protein T_f estimations with sigmoid function fits as a function of length-normalized coevolutionary energy of the respective sequences. (*B*) Cooperativity score ρ is shown in a color scale on a plane defined by the average normalized internal energy difference $\langle |e_j^i - e_k^i| \rangle$ and the average nonzero normalized surface energies $-\langle e^s \rangle$. Level curves (gray) were obtained with a polynomial fit.

elements (normalizing by element length) were compensated by the corresponding surface energies ϵ^s_{jk}, elements can fold cooperatively in an all-or-none transition. This would be the case of a full consensus protein with exactly duplicated repeats. But complexity arises because this is not the case for natural proteins, as one can see in Fig. 5*B*. Each point corresponds to a sequence of the selected set on a plane, defined by the average normalized internal energy difference $\langle|e^i_j - e^i_k|\rangle$ and the average nonzero normalized surface energies $-\langle e^s\rangle$, while the cooperativity score ρ is represented by the color scale. Cooperativity smoothly changed from two-state proteins with strong surface terms and low heterogeneity to full downhill proteins on the opposite corner of the plot. In the central region, sequences could present, for instance, a sharp transition and a pretransition as p16 and/or many barriers and intermediaries.

Interestingly, there are no arrays that are simultaneously highly heterogeneous and strongly coupled, so the top right region of the plot is empty (Fig. 5*B*). This is not a particular feature of the folding model, but a characteristic of the structure of the sequence space that is captured by the evolutionary model. It is possible that there are evolutionary constraints that impose a limit on the coupling strengths between energetically dissimilar sequence fragments. We fitted a polynomial function of the energetic heterogeneity $\langle|e^i_j - e^i_k|\rangle$ and interaction average $\langle e^s\rangle$ to define a phase diagram (Fig. 5*B*) and to predict cooperativity ρ directly from the amino acid sequence.

The evolutionary model can be used to generate sequence ensembles with a Monte Carlo run, such that they reproduced the natural ANK dataset features (details are in *SI Appendix*). For a generated ensemble of 4,000 sequences with the same protein-length distribution of the selected set, T_f and ρ trends with energy roughly held (*SI Appendix*, Fig. S15). Hence, it is possible to generate a large ensemble of sequences and then select a subset with the desired cooperativity and stability, without computing the Ising model.

Concluding Remarks

Although superficially simple, the energy landscapes of repeat proteins show very rich behaviors. We explored here the folding of thousands of naturally occurring Ankyrin repeat proteins and found that a multiplicity of mechanisms can be coded in sequences with a common low-temperature fold. On the one hand, fully cooperative all-or-none transition is obtained when the proteins are composed by sequence-similar elements and strong interactions between them. On the other hand, noncooperative element-by-element intermittent folding becomes the rule when the elements are dissimilar and the interactions between them are energetically weak. In between these extremes, cooperative folding domains may emerge. Along the dataset, 73% of elements fold together with other ones, forming apparent domains. Notably, we found that there is not a characteristic domain size and capture a scale-free domain formation. Particularly, the first domain to emerge covers about half of the repeat array, and the rest of the chain folds upon it.

We want to make it clear that, at this point, the main part of the results presented in this work are predictions of folding mechanisms. It remains to be experimentally tested to which extent these can be modulated by the sequences as we propose. In principle, we can point out which natural proteins are predicted to have distinctive folding dynamics and could be synthesized and purified, such that their equilibrium folding properties can be measured and contrasted to the predictions, as done for some ANK proteins, which we used to test the initial model (*SI Appendix*, Fig. S9). All the data for natural proteins and the code to simulate sequences are deposited in GitHub (*Materials and Methods*).

We propose that the overall behavior of the proteins can be projected in two dimensions that capture the cooperativity and the stability of the arrays. Using purely evolutionary information, it is possible to predict the thermodynamics of these systems, beyond the effect of single amino acid substitutions, but the truly collective phenomena of protein folding. Strikingly, there are no examples of natural proteins for which the sequence heterogeneity is high and the interactions between them are strong. This effect can be attributed to the internal structure of the learned evolutionary field, as simulated sequences show the same distributions.

Recent works have shown that it is possible to use coevolutionary information in the design of enzymes (48) and modular repressors (49). Here, we obtained variegated folding mechanisms with simulated sequences, which, in principle, can be used to design repeat proteins with desired folding properties and mechanical functions—for example, in their nano-spring behavior (50).

The biological function of most Ankyrin repeat arrays is thought to be mediated by specific protein–protein interactions, and, for many of them, the folding of repeats is coupled to the binding of their targets. We identified examples for which the calculations match the known experimental region that undergoes transitions. We speculate that the rich folding behavior we identified here can be related to the biological function of these proteins—for example, in the identification of the regions that undergo transitions at low temperatures as binding or allosteric regions. Further experiments could discriminate how widely distributed these effects are.

Furthermore, the general model we implemented could be transferable to other systems, those for which it is reasonable to separate the sequence in fragments, each of which folds as a single cooperative module to assign the folding model and for systems where enough sequences are available to learn the evolutionary local fields and couplings. Despite the recent interest on the success of structure prediction tools (51), we should bear in mind that the dynamics of natural proteins is fundamental to their biological activity and evolution. The occupation of excited states on the energy landscapes is crucial in determining the interactions that proteins juggle in the crowded interior of cells. We presented here a way to model these dynamics solely from sequence information, which may well be applied to other types of proteins.

Materials and Methods

Sequence Data Curation. We used subsets of 1.2 million Ankyrin repeat sequence alignments, previously built and characterized (32). More details are provided in *SI Appendix, SI Methods*.

Evolutionary Model for Repeat Arrays. We used the evolutionary energy fields learned with a statistical model from an ANK multiple-sequence alignment (MSA), which included internal, but not terminal, repeats from the same full dataset previously described (32) and arrays up to 40 repeats long. Briefly, the model combines DCA and an explicit evolution mechanism of duplications and deletions of repeats. The maximum-entropy model uses as constraints empirical MSA single-site and two-point amino acid frequencies, pairwise repeat identity, and array length distribution.

Statistical inference was made with a Boltzmann Learning algorithm, obtaining energy fields $\tilde{h}_a(\sigma_a)$ and $\tilde{J}_{ab}(\sigma_a, \sigma_b)$. More details about the model definition, parameter inference, and validation are provided in *SI Appendix*.

Ising Model Elements Assignment. The multiple-repeat sequence alignment we worked with has the amino acid pattern TPLH on positions 10 to 13. We divided each repeat in two fragments using secondary and tertiary structural

information reported in the literature (52). We labeled residues 1 to 18 containing β-hairpins and an α-helix as the fragment *A* and residues 19 to 33 containing the second α-helix as the fragment *B*. Given that the dataset only contained concatenated 33 amino acid repeats, any repeat array can be mapped to a periodic succession of *A*–*B* elements. Repeats with less than 33 residues include gaps ("−") to complete all the 33 positions. Although the evolutionary field was learned, including gaps in the alphabet, and rigorously, there is a folding energy contribution to be considered, we set both energetic and entropic gap positions contribution to zero.

Ising Model Implementation. We performed Monte Carlo Metropolis algorithm simulations of the finite Ising model with a python routine. The code is available at GitHub (https://github.com/eagalpern/folding-ising). Simulation total time, transient time, and equilibration time parameters scale linearly with protein length and were obtained with an autocorrelation analysis. Deletions, unknown amino acids, and missing residues at the beginning or ending of sequences were excluded from all calculations. For the selected dataset, simulations were made for 500 equispaced temperatures in an interval, such that the system folded and unfolded completely.

Free-Energy Profiles Approximation. We obtained free-energy profiles approximating the probability of states *s* with *Q* folded elements with the Metropolis Monte Carlo sampling. We considered together sampled states for simulations performed in a window of the 10 closest temperatures. The profiles we used are computed as

$$\Delta f(Q) = -kT\, log\left(\frac{\sum_{s|Q} N(s)}{\sum_{s} N(s)}\right), \qquad [4]$$

where T is the average temperature, $N(s)$ are the counts of state s, and $s|Q$ are the states with Q folded elements.

Apparent Domains. Elements j and k were assigned to the same domain if $|T_f^j - T_f^k| < 5$, where the folding temperature T_f^j was obtained by a sigmoid fit of the folding probability of element j. Domain folding temperature is the average $\langle T_f^j \rangle$ for j belonging the domain. Overlapping domains were separated into the minimum number of nonoverlapping ones. If more than one separation was possible, temperature difference between domains were maximized.

Folding Temperature. To fit folding temperatures T_f, we approximated the fraction folded $m(T)$ as

$$m(T) = \frac{m_{max}}{1 + e^{a(T - T_f)}}, \qquad [5]$$

where $m_{max} \in [0, 1]$. We used scipy library curve_fit to fit and get σ_{T_f}, which we used as T_f errors.

Cooperativity Phase Diagram Fit. We made multivariate polynomial fits on the selected set, making fivefold cross-validation and using python sklearn library preprocessing.PolynomialFeatures and linear_model.LinearRegression. We compared performance measured by predicted ρ rms error for 1- to 9-degree polynomial and kept the best one, 3-degree.

Data Availability. Code and data have been deposited in GitHub (https://github.com/eagalpern/folding-ising) (53).

ACKNOWLEDGMENTS. This work used computational resources from Universidad Nacional de Córdoba (https://ccad.unc.edu.ar/), which are part of El Sistema Nacional de Computación de Alto Desempeño–Ministerio de Ciencia, Tecnología e Innovación, República Argentina. This work was supported by the Consejo de Investigaciones Científicas y Técnicas; the Agencia Nacional de Promoción Científica y Tecnológica (PICT2016-1467 [to D.U.F.]); and Universidad de Buenos Aires (UBACYT 2018 20020170100540BA). Additional support was from the NASA Astrobiology Institute and Grant 80NSSC18M0093 Proposal: ENIGMA: EVOLUTION OF NANOMACHINES IN GEOSPHERES AND MICROBIAL ANCESTORS (NASA ASTROBIOLOGY INSTITUTE CYCLE 8); and European Research Council Consolidator Grant 724208. We thank I. E. Sánchez and P. G. Wolynes for stimulating discussions and comments during the development of this work.

Author affiliations: [a]Protein Physiology Lab, Facultad de Ciencias Exactas y Naturales, Departamento de Química Biológica, Universidad de Buenos Aires, C1428EGA Buenos Aires, Argentina; [b]Consejo Nacional de Investigaciones Científicas y Técnicas–Universidad de Buenos Aires, Instituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales, C1428EGA Buenos Aires, Argentina; and [c]Laboratoire de Physique de l'École Normale Supérieure, CNRS, Paris Sciences et Lettres University, Sorbonne Université, and Université de Paris-Cité, 75005 Paris, France

1. J. D. Bryngelson, P. G. Wolynes, Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7524–7528 (1987).
2. P. G. Wolynes, Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie* **119**, 218–230 (2015).
3. D. U. Ferreiro, E. A. Komives, P. G. Wolynes, Frustration, function and folding. *Curr. Opin. Struct. Biol.* **48**, 68–73 (2018).
4. L. Paladin *et al.*, RepeatsDB in 2021: Improved data and extended classification for protein tandem repeat structures. *Nucleic Acids Res.* **49** (D1), D452–D457 (2021).
5. R. Espada *et al.*, Repeat proteins challenge the concept of structural domains. *Biochem. Soc. Trans.* **43**, 844–849 (2015).
6. M. Petersen, D. Barrick, Analysis of tandem repeat protein folding using nearest-neighbor models. *Annu. Rev. Biophys.* **50**, 245–265 (2021).
7. T. Aksel, D. Barrick, Direct observation of parallel folding pathways revealed using a symmetric repeat protein system. *Biophys. J.* **107**, 220–232 (2014).
8. K. W. Tripp, D. Barrick, Rerouting the folding pathway of the Notch ankyrin domain by reshaping the energy landscape. *J. Am. Chem. Soc.* **130**, 5681–5688 (2008).
9. D. U. Ferreiro, E. A. Komives, The plastic landscape of repeat proteins. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7735–7736 (2007).
10. A. Schüler, E. Bornberg-Bauer, Evolution of protein domain repeats in metazoa. *Mol. Biol. Evol.* **33**, 3170–3182 (2016).
11. A. Kumar, J. Balbach, Folding and stability of ankyrin repeats control biological protein function. *Biomolecules* **11**, 840 (2021).
12. C. C. Mello, D. Barrick, An experimentally determined protein folding energy landscape. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 14102–14107 (2004).
13. R. Espada, R. G. Parra, T. Mora, A. M. Walczak, D. U. Ferreiro, Inferring repeat-protein energetics from evolutionary information. *PLOS Comput. Biol.* **13**, e1005584 (2017).
14. D. U. Ferreiro, A. M. Walczak, E. A. Komives, P. G. Wolynes, The energy landscapes of repeat-containing proteins: Topology, cooperativity, and the folding funnels of one-dimensional architectures. *PLOS Comput. Biol.* **4**, e1000070 (2008).
15. R. G. Parra, R. Espada, N. Verstraete, D. U. Ferreiro, Structural and energetic characterization of the ankyrin repeat protein family. *PLOS Comput. Biol.* **11**, e1004659 (2015).
16. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 67–72 (2009).
17. F. Morcos *et al.*, Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293–E1301 (2011).
18. J. A. de la Paz, C. M. Nartey, M. Yuvaraj, F. Morcos, Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 5873–5882 (2020).
19. A. Contini, G. Tiana, A many-body term improves the accuracy of effective potentials based on protein coevolutionary data. *J. Chem. Phys.* **143**, 07B608_1 (2015).
20. M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, M. Weigt, Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1. *Mol. Biol. Evol.* **33**, 268–280 (2016).
21. A. Haldane, W. F. Flynn, P. He, R. S. Vijayan, R. M. Levy, Structural propensities of kinase family proteins from a Potts model of residue co-variation. *Protein Sci.* **25**, 1378–1384 (2016).
22. K. S. Tang, B. J. Guralnick, W. K. Wang, A. R. Fersht, L. S. Itzhaki, Stability and folding of the tumour suppressor protein p16. *J. Mol. Biol.* **285**, 1869–1886 (1999).
23. Y. Guo *et al.*, Contributions of conserved TPLH tetrapeptides to the conformational stability of ankyrin repeat proteins. *J. Mol. Biol.* **399**, 168–181 (2010).
24. K. S. Tang, A. R. Fersht, L. S. Itzhaki, Sequential unfolding of ankyrin repeats in tumor suppressor p16. *Structure* **11**, 67–73 (2003).
25. D. U. Ferreiro *et al.*, Stabilizing $i\kappa b\alpha$ by "consensus" design. *J. Mol. Biol.* **365**, 1201–1216 (2007).
26. I. DeVries, D. U. Ferreiro, I. E. Sánchez, E. A. Komives, Folding kinetics of the cooperatively folded subdomain of the $i\kappa b\alpha$ ankyrin repeat domain. *J. Mol. Biol.* **408**, 163–176 (2011).
27. T. O. Street, C. M. Bradley, D. Barrick, An improved experimental system for determining small folding entropy changes resulting from proline to alanine substitutions. *Protein Sci.* **14**, 2429–2435 (2005).
28. N. P. Schafer *et al.*, Discrete kinetic models from funneled energy landscape simulations. *PLoS One* **7**, e50635 (2012).
29. I. J. L. Byeon *et al.*, Tumor suppressor p16INK4A: Determination of solution structure and analyses of its interaction with cyclin-dependent kinase 4. *Mol. Cell* **1**, 421–431 (1998).
30. C. Yuan, T. L. Selby, J. Li, I. J. L. Byeon, M. D. Tsai, Tumor suppressor INK4: Refinement of p16INK4A structure and determination of p15INK4B structure by comparative modeling and NMR data. *Protein Sci.* **9**, 1120–1128 (2000).
31. G. Interlandi, G. Settanni, A. Caflisch, Unfolding transition state and intermediates of the tumor suppressor p16INK4a investigated by molecular dynamics simulations. *Proteins* **64**, 178–192 (2006).
32. E. A. Galpern, M. I. Freiberger, D. U. Ferreiro, Large Ankyrin repeat proteins are formed with similar and energetically favorable units. *PLoS One* **15**, e0233865 (2020).
33. H. Inada, E. Procko, M. Sotomayor, R. Gaudet, Structural and biochemical consequences of disease-causing mutations in the ankyrin repeat domain of the human TRPV4 channel. *Biochemistry* **51**, 6195–6206 (2012).
34. Q. Yang, H. Liu, Z. Li, Y. Wang, W. Liu, Purification and mutagenesis studies of TANC1 ankyrin repeats domain provide clues to understand mis-sense variants from diseases. *Biochem. Biophys. Res. Commun.* **514**, 358–364 (2019).

35. W. Pan *et al.*, Structural insights into ankyrin repeat-mediated recognition of the kinesin motor protein KIF21A by KANK1, a scaffold protein in focal adhesion. *J. Biol. Chem.* **293**, 1944–1956 (2018).
36. M. C. Baxa, E. J. Haddadian, J. M. Jumper, K. F. Freed, T. R. Sosnick, Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 15396–15401 (2014).
37. J. A. D'Aquino *et al.*, The magnitude of the backbone conformational entropy change in protein folding. *Proteins* **25**, 143–156 (1996).
38. G. I. Makhatadze, P. L. Privalov, On the entropy of protein folding. *Protein Sci.* **5**, 507–510 (1996).
39. D. U. Ferreiro, E. A. Komives, Molecular mechanisms of system control of nf-κb signaling by iκbα. *Biochemistry* **49**, 1560–1567 (2010).
40. C. M. Bradley, D. Barrick, Limits of cooperativity in a structurally modular protein: Response of the Notch ankyrin domain to analogous alanine substitutions in each repeat. *J. Mol. Biol.* **324**, 373–386 (2002).
41. D. Barrick, D. U. Ferreiro, E. A. Komives, Folding landscapes of ankyrin repeat proteins: Experiments meet theory. *Curr. Opin. Struct. Biol.* **18**, 27–34 (2008).
42. N. D. Werbeck, L. S. Itzhaki, Probing a moving target with a plastic unfolding intermediate of an ankyrin-repeat protein. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7863–7868 (2007).
43. D. U. Ferreiro, P. G. Wolynes, The capillarity picture and the kinetics of one-dimensional protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 9853–9854 (2008).
44. A. Kohl *et al.*, Designed to be stable: Crystal structure of a consensus ankyrin repeat protein. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 1700–1705 (2003).
45. S. K. Wetzel, G. Settanni, M. Kenig, H. K. Binz, A. Plückthun, Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *J. Mol. Biol.* **376**, 241–257 (2008).
46. E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, A. Bairoch, UniProtKB/Swiss-Prot. *Methods Mol. Biol.* **406**, 89–112 (2007).
47. P. Tian, J. M. Louis, J. L. Baber, A. Aniana, R. B. Best, Co-evolutionary fitness landscapes for sequence design. *Angew. Chem. Int. Ed. Engl.* **57**, 5674–5678 (2018).
48. W. P. Russ *et al.*, An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).
49. X. L. Jiang, R. P. Dimas, C. T. Y. Chan, F. Morcos, Coevolutionary methods enable robust design of modular repressors by reestablishing intra-protein interactions. *Nat. Commun.* **12**, 5592 (2021).
50. G. Lee *et al.*, Nanospring behaviour of ankyrin repeats. *Nature* **440**, 246–249 (2006).
51. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
52. S. G. Sedgwick, S. J. Smerdon, The ankyrin repeat: A diversity of interactions on a common structural framework. *Trends Biochem. Sci.* **24**, 311–316 (1999).
53. E. A. Galpern, Folding Ising, code and data for "Evolution and folding of repeat proteins." GitHub. https://github.com/eagalpern/folding-ising. Deposited 3 June 2022.