Broken TALEs: Transcription Activator-like Effectors Populate Partly Folded States

Kathryn Geiger-Schuller[1,2] and Doug Barrick[2,*]

[1]Program in Molecular Biophysics and [2]T.C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, Maryland

ABSTRACT Transcription activator-like effector proteins (TALEs) contain large numbers of repeats that bind double-stranded DNA, wrapping around DNA to form a continuous superhelix. Since unbound TALEs retain superhelical structure, it seems likely that DNA binding requires a significant structural distortion or partial unfolding. In this study, we use nearest-neighbor "Ising" analysis of consensus TALE (cTALE) repeat unfolding to quantify intrinsic folding free energies, coupling energies between repeats, and the free energy distribution of partly unfolded states, and to determine how those energies depend on the sequence that determines DNA-specificity (called the "RVD"). We find a moderate level of cooperativity for both the HD and NS RVD sequences (stabilizing interfaces combined with unstable repeats), as has been seen in other linear repeat proteins. Surprisingly, RVD sequence identity influences both the overall stability and the balance of intrinsic repeat stability and interfacial coupling energy. Using parameters from the Ising analysis, we have analyzed the distribution of partly folded states as a function of cTALE length and RVD sequence. We find partly unfolded states where one or more repeats are unfolded to be energetically accessible. Mixing repeats with different RVD sequences increases the population of partially folded states. Local folding free energies plateau for central repeats, suggesting that TALEs access partially folded states where a single internal repeat is unfolded while adjacent repeats remain folded. This breakage should allow TALEs to access superhelically-broken states, and may facilitate DNA binding.

INTRODUCTION

Transcription activator-like effectors (TALEs) are bacterial proteins containing a domain of tandem 34-residue repeats that binds to specific DNA sequences and affect transcription of host genes (1,2). TALE repeat domains have an average of 17.5 repeats, although this number varies (3). Repeats have high sequence identity, with most of the variability at repeat positions 12 and 13. These two residues, together called repeat variable diresidues (RVDs), impart DNA binding specificity, and identities at positions 12 and 13 can be used to design TALE proteins that bind specific DNA sequences (4–6). Using this specificity code, TALE nucleases (TALENs) have been engineered for genome editing purposes (7,8). TALE repeats have also been used to design proteins that bind to DNA and activate or repress transcription (9–13), modify DNA by demethylation (14), and probe chromatin dynamics by encoding fluorophores with high sequence specificity (15,16). One challenge to TALE-based genome

editing is cloning difficulties resulting from the large number of repeats needed for affinity/specificity (17). Enhancing the affinity of TALE repeats would thus be beneficial for genome editing and molecular genetic studies, allowing high-specificity target recognition with fewer repeats.

TALEs in their unbound state are superhelical, with 11 repeats per turn (18) and are not likely to thread onto DNA easily. It seems likely that a conformational change is required for binding. One way to bring about such a change is a global deformation of the native state. Some degree of native-state plasticity is suggested from comparison of the free and bound states (Fig. S1 in Supporting Material) but does not seem to be enough to adequately open the structure. Alternatively, a binding competent state could be reached by a localized structural transition, either by disrupting one or more interfaces between repeats, and or by unfolding one or more repeats.

Ising analysis provides a direct means to determine the cooperativity of linear arrays of identical sequence repeats, and for quantifying the populations of partly folded states (19–22). By studying the length dependence of stability, one-dimensional (1D) Ising analysis can resolve

the energetics of folding individual repeats from the energetics of formation of interfaces between repeats. Two types of repeat proteins have been subjected to Ising analysis (TPR (20,23) and ankyrin repeats (19,21,22)). By applying 1D Ising analysis to identical TALE repeat arrays of different lengths, we can resolve folding free energies into intrinsic and interfacial components, quantify the extent of cooperativity in folding, and determine the populations of partially folded states that may facilitate DNA binding.

Here, we characterize the equilibrium stability of a series of TALE constructs with varying length and RVD sequence using nearest-neighbor Ising analysis. The length dependence and Ising analysis demonstrate an intermediate level of coupling between repeats. Local folding free energies (ΔG°_{local}), calculated from intrinsic and interfacial free energies, suggest significant populations of partially unfolded states. Surprisingly, the extent of coupling and local folding free energy profiles depend on the sequence of the RVDs. This dependence leads to a stability switch between NS- and HD-containing TALE arrays at a length of eight repeats. The stabilities of mixed NS and HD constructs demonstrate that RVD sequence identity partitions asymmetrically into its N- and C-terminal interfaces, introducing further variation in local folding free energies.

MATERIALS AND METHODS

Cloning, expression, and purification

Consensus TALE repeat constructs were cloned with C-terminal His$_6$ tags via an in-house version of Golden Gate cloning (24). TALE constructs were grown in BL21(T1R) cells at 37°C to an optical density of 0.6–0.8 and induced with 1 mM isopropyl-β-D-1 thiogalactopyranoside. Following cell pelleting, resuspension, and lysis, proteins were purified by resuspending the insoluble material in 6M urea, 300 mM NaCl, and 10 mM Tris (pH 7.4). Constructs were loaded onto an Ni-NTA column. Protein was eluted using 250 mM imidazole and refolded during dialysis into 300 mM NaCl, 5% glycerol, and 10 mM Tris (pH 7.4).

Circular dichroism spectroscopy

Circular dichroism (CD) measurements were collected using an AVIV model 400 CD Spectrometer (Aviv Associates, Lakewood, NJ). Far-ultraviolet (UV) CD scans were collected at 25°C using an 0.1 cm pathlength quartz cuvette, with protein concentrations of 15–30 μM. Buffer scans were recorded and were subtracted from the raw CD data.

Urea-induced unfolding transitions

For short constructs, CD-monitored unfolding titrations at 222 nm were generated with an automated titrator. For N-capped constructs six repeats or longer, and all constructs seven repeats or longer, slow relaxation kinetics prevented us from collecting automated titrations; thus, we performed manual urea titrations for these constructs. Solutions containing 0 and 8 M urea, each with 2 μM protein, were combined in various proportions using Hamilton syringes. Samples equilibrated for 12–24 h at room temperature before monitoring CD signal at 222 nm.

Ising analysis

To determine the intrinsic and interfacial free energies for folding of cTALE arrays, and to analyze energies and populations of partly folded states, we used a 1D Ising formalism (25,26). In this model, intrinsic folding and interfacial interaction between nearest neighbors are represented using equilibrium constants κ and τ, respectively, as follows:

$$\kappa_N = e^{-(\Delta G_N - 4m[urea])/RT}, \tag{1.1}$$

$$\kappa_R = e^{-(\Delta G_R - m[urea])/RT}, \tag{1.2}$$

$$\kappa_C = e^{-(\Delta G_C - m[urea])/RT}, \tag{1.3}$$

$$\tau_N = e^{-\left(\Delta G_{N,\ i+1}\right)/RT}, \tag{1.4}$$

$$\tau_R = e^{-\left(\Delta G_{R,\ i+1}\right)/RT}. \tag{1.5}$$

In the current analysis, the intrinsic folding free energies of N (solubilizing N-terminal cap), R (consensus repeat), and C (solubilizing C-terminal cap) are each considered to be unique. The interfacial interactions of the R:R and R:C pairs are considered to be identical, but that the N:R pair is considered to be unique. Denaturant dependences are built into the intrinsic (but not the interfacial) terms. The urea dependence of the N-terminal cap is scaled by a factor of four because there are four TALE-like repeats in the N-terminal cap.

Using these equilibrium constants, a partition function q can be constructed for an n-repeat construct by multiplying two-by-two transfer matrices as follows:

$$q = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_N \tau_N & \kappa_N \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_R \tau_R & \kappa_R \\ 1 & 1 \end{bmatrix}^{n-2} \begin{bmatrix} \kappa_C \tau_R & \kappa_C \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{2}$$

This representation differs from several previous expressions in that the matrices in Eq. 2 correlate to the next repeat (rather than the previous repeat) (25). Though this rephrasing does not alter q, it associates the κ and τ terms for the same repeat. The fraction folded (f_{folded}) can be determined as follows:

$$f_{folded} = \frac{1}{nq} \left(\kappa_N \frac{\partial q}{\partial \kappa_N} + \kappa_R \frac{\partial q}{\partial \kappa_R} + \kappa_C \frac{\partial q}{\partial \kappa_C} \right). \tag{3}$$

A specific example for a three-repeat cTALE construct is given in the Supporting Material and Methods (Fig. S2).

Ising parameters were determined by nonlinear least squares using an in-house Python program (written by J. Marold) (23) by globally fitting Eq. 3 to urea-induced unfolding transitions. Confidence intervals (CI) were determined by performing 2000 bootstrap iterations (95%).

Calculation of local folding free energies within TALE arrays

Local folding free energies, ΔG°_{local}, are calculated using fitted Ising parameters in Table 1 (see also Supporting Materials and Methods). Summing all statistical weights for states where the ith repeat is folded and dividing by the statistical weights where the ith repeat is unfolded gives a local

TABLE 1   Summary of Thermodynamic Parameters Obtained from Ising-Fit

Intrinsic Terms	ΔG_N	ΔG_{NS}	ΔG_{HD}	ΔG_C	m_i	
Best-Fit	−4.42	5.89	3.49	7.14	−0.50	
95% CI[a]	−4.80, −4.03	5.47, 6.33	3.13, 3.82	6.78, 7.50	−0.52, −0.48	
Interfacial Terms	$\Delta G_{N,\,i+1}$	$\Delta G_{NS,\,NS+1}$	$\Delta G_{HD,\,HD+1}$	$\Delta G_{C,\,HD-1}$	$\Delta G_{NS,\,HD+1}$	$\Delta G_{HD,\,NS+1}$
Best-Fit	−0.85	−7.79	−5.02	−8.49	−7.58	−4.73
95% CI[a]	−0.94, −0.76	−8.30, −7.30	−5.42, −4.59	−9.00, −8.00	−8.09, −7.09	−5.24, −4.19

All values obtained are kcal/mol with the exception of m_i, which is kcal/mol/M.

[a]95% confidence intervals are from 2000 iterations of bootstrap analysis.

equilibrium constant for folding (K_{local}). The local free energy of folding is calculated using the following formula:

$$\Delta G_{local} = -RT \ln K_{local}. \qquad (4)$$

RESULTS

Consensus TALE design

To design a consensus TALE (cTALE) repeat sequence for Ising analysis, the Hidden Markov model search tool was used to collect and align 3,667 TALE repeat sequences (27). From this alignment, Skylign was used to create an HMM logo (Fig. 1 A), where the height of each residue at each position is proportional to its conservation (28). The most-conserved residue at each position was selected for the consensus-based TALE sequence except at position 30, where arginine was chosen instead of cysteine to simplify folding studies. In addition, the two RVD residues, positions 12 and 13, were varied to generate two common recognition sequences (NS and HD; Fig. 1 B). HD is the most-common RVD sequence, and thus conforms to the consensus sequence. In contrast, NS is less frequent, and provides an interesting point of comparison, both for stability studies and in future work, DNA binding.

We initially built consensus TALE arrays without terminal capping repeats, but found these constructs to self-associate by sedimentation velocity analytical ultracentrifugation (AUC; data not shown). Previous studies with ankyrin repeat and TPR consensus constructs have shown that polar N- and C-terminal caps are essential for solubility (19,20,22,23,29,30). Thus, we designed N- and C-terminal caps to help solubilize our consensus TALE arrays.

For the N-terminal cap, we selected a conserved 149 residue N-terminal extension of the naturally occurring TALE gene product PthXo1. Crystal structures show that this N-terminal extension forms four cryptic repeats, with similar structure to TALE repeats despite significant sequence differences from the TALE consensus (Fig. 1 C) (31,32). This N-cap has been shown to be resistant to proteolysis and is required for full transcription activation (31). The C-terminal cap was designed by changing consensus hydrophobic residues predicted to be solvent exposed to polar or charged residues (Fig. 1 D). Sedimentation velocity AUC experiments demonstrated that constructs with both the N- and C-caps are monomeric (Fig. S3). Subsequent AUC experiments showed singly capped (either N- or C-terminal) constructs are also soluble and monomeric. Including these singly capped constructs allows us to resolve the thermodynamic contributions of the capping repeats from the internal repeats.
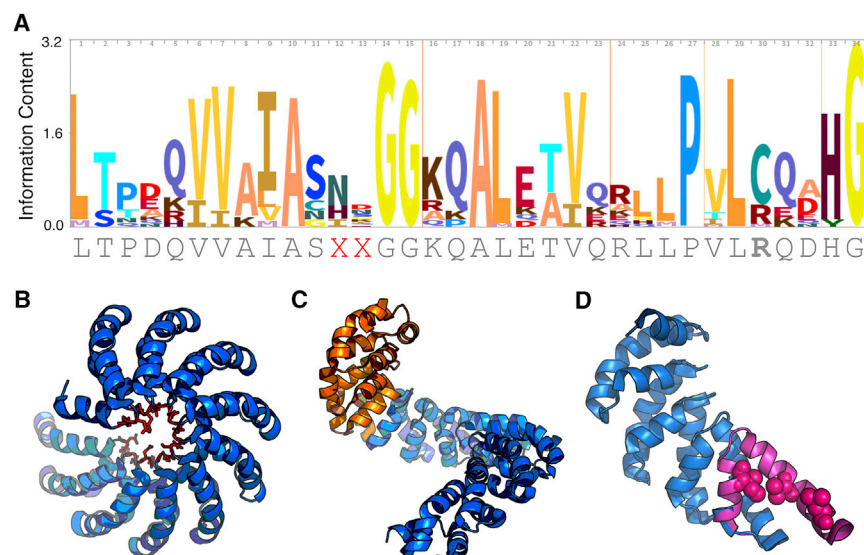

FIGURE 1   Sequence conservation and structure of TALE repeats. (A) A sequence logo of TALE repeats showing conservation at each of the 34 positions. The sequence below is the consensus sequence used in these studies. At the RVD positions (12,13), two common recognition motifs were selected (NS and HD). (B) Crystal structure of dHax3 highlighting location of RVDs (red sticks) (18). (C) The N-terminal cap (orange) is a conserved extension of the repeat domain composed of four TALE-like repeats (31). (D) The C-terminal cap was designed by substituting solvent exposed hydrophobic residues to polar or charged residues (pink spheres). To see this figure in color, go online.

To confirm that cTALE repeats have α-helical secondary structure, far-UV CD spectra were collected for various cTALE constructs. The spectra of all constructs are consistent with α-helical structure and are similar in shape to the far-UV CD spectrum of the naturally occurring PthXo1 TALE domain (Figs. 2 and S4). Helical structure is retained when either the N- or C-cap is absent. Also, consensus TALEs retain DNA-binding activity (Fig. S5).

The RVD sequence affects cTALE stability

To determine the effect of RVD sequence on stability, urea-induced unfolding transitions were measured by CD spectroscopy for a construct with NS RVDs ((NS)$_6$C, with six NS repeats and a C-cap) and an otherwise identical construct with HD repeats ((HD)$_6$C, Fig. 3). Both unfolding transitions are sigmoidal and are well-fitted with a two-state model. The unfolding transitions of the HD and NS constructs have similar slopes (and thus, similar m-values), but have significantly different unfolding midpoints (C_m values). As a result, the NS and HD constructs have different free energies of unfolding ($\Delta G^{\circ}_{H_2O}$). The sigmoidal transitions and high m-values are consistent with a high level of cooperativity in unfolding, suggesting strong interfacial coupling between repeats. The differences in $\Delta G^{\circ}_{H_2O}$ values indicate that RVD identity affects intrinsic folding energy, interrepeat coupling energy, or both.

Length and capping dependence of NS TALE stability

To resolve the intrinsic stability from the interfacial coupling energy between TALE repeats, urea-induced unfolding transitions were measured for constructs with different numbers of
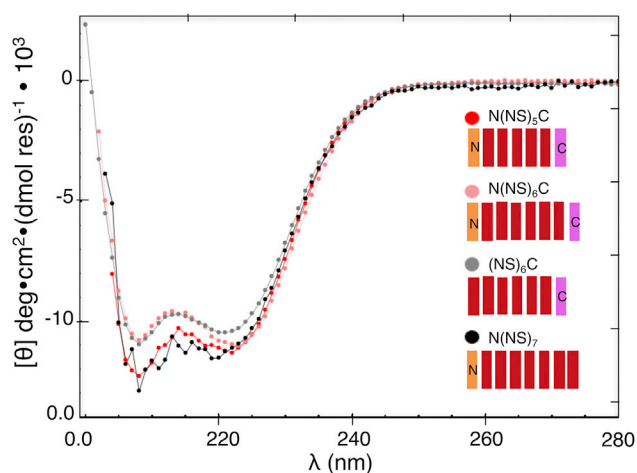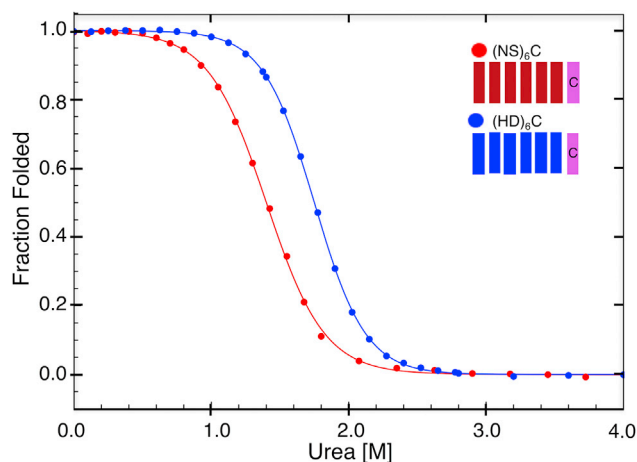

FIGURE 3 Consensus TALE stability is dependent on RVD sequence. Urea-induced unfolding transitions of TALE constructs show cooperative, two-state unfolding transitions. Urea-induced unfolding transitions of each construct were fitted with a two-state model for unfolding (*solid lines*). Global stabilities, based on unfolding midpoints, vary significantly with RVD sequence, although slopes of the transitions do not. Conditions: 300 mM NaCl, 10 mM Tris HCl (pH 7.4), 5% glycerol, 25°C. To see this figure in color, go online.

NS repeats. Because the number of repeats and interfaces in each construct differ, analyzing the unfolding of constructs of different lengths allows the intrinsic and interfacial energies to be treated as independent variables (26). To account for sequence differences in the N- and C-terminal repeat sequences, we included constructs lacking either the N- or the C-terminal cap. These constructs are crucial for untangling intrinsic and interfacial free energies from variations due to capping substitutions.

For a given capping structure, adding internal NS repeats increases stability (compare N(NS)$_5$C and N(NS)$_6$C, as well as N(NS)$_6$, N(NS)$_7$, and N(NS)$_8$; Fig. 4 A), again consistent with strong energetic coupling between repeats. For constructs with six internal NS repeats, the construct containing both the N- and C-terminal cap has the highest midpoint, followed by the construct with only the N-terminal cap (Fig. 4 B). The construct with only the C-terminal cap has the lowest midpoint.

Transitions for N-capped constructs are not as steep as for constructs lacking N-caps. This suggests that the unfolding transitions of N-capped constructs are not two-state. It seems unlikely that adding the N-terminal cap uncouples the (NS)$_6$C unfolding transition; rather, the decreased slope for the N-capped constructs suggests weak coupling between the N-cap and central repeats combined with a high intrinsic stability for the N-term capping segment.

Length and capping dependence of HD TALE stability

To better understand the origins of the stability differences among different RVD sequences (Fig. 3), urea-induced

FIGURE 2 Doubly- and singly-capped TALE consensus constructs are α-helical. Far-UV CD spectra of TALE NS repeat constructs with N-cap, C-cap, and both caps. Spectra are consistent with folded, α-helical structures. Conditions: 300 mM NaCl, 10 mM Tris HCl (pH 7.4), 5% glycerol, 25°C. To see this figure in color, go online.
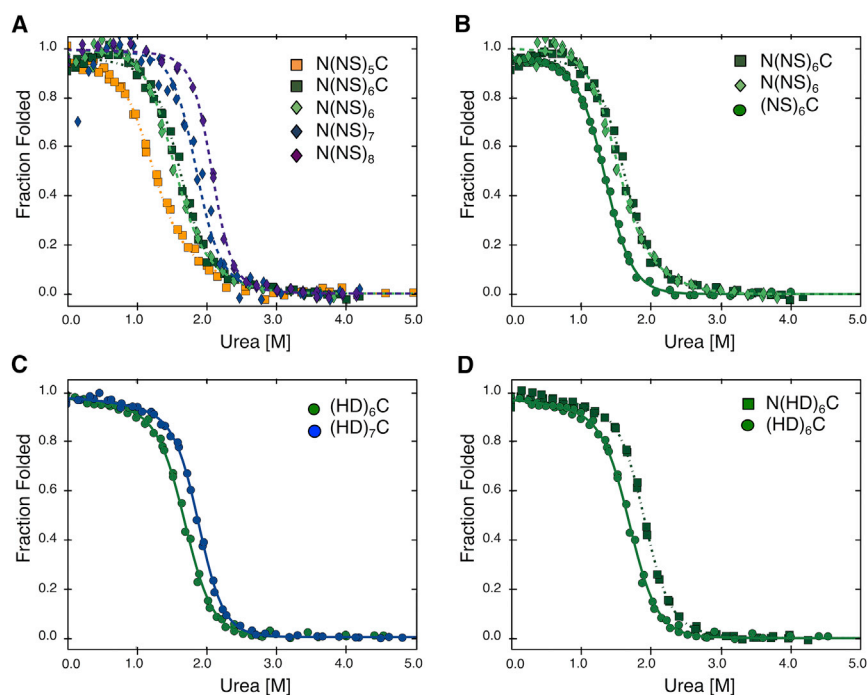
FIGURE 4 Length- and capping-dependence of TALE HD- and NS-RVD stability. Urea-induced unfolding transitions of TALE constructs were globally fitted using a heterogeneous nearest-neighbor Ising model (N-capped, *dashed lines*; C-capped, *solid lines*; doubly capped, *dotted lines*). (*A*) NS-type repeat constructs with increasing repeat number, for N-capped and doubly capped constructs. Stability increases with number of repeats are shown. (*B*) Constructs with six NS-type repeats with varying capping identities are shown. $N(NS)_6C$ is most stable followed by $N(NS)_6$. $(NS)_6C$ has the smallest C_m but the largest slope. (*C*) HD-type repeat constructs with increasing repeat number are shown. The increase in midpoint of the transition between $(HD)_6C$ and $(HD)_7C$ is less than the increase in midpoint between $N(NS)_5C$ and $N(NS)_6C$ (*A*). (*D*) Constructs with six HD-type repeats with varying capping identities are shown. As with the NS-type repeats (*B*), the doubly capped construct is more stable than the singly capped construct. Conditions: 300 mM NaCl, 10 mM Tris HCl (pH 7.4), 5% glycerol, 25°C. To see this figure in color, go online.

unfolding transitions were measured for HD-type repeats of various lengths and capping structures. Adding an internal HD repeat is stabilizing, as shown in Fig. 4 *C*. Although the C_m values increase with repeat number for the HD series (compare $(HD)_6C$ and $(HD)_7C$ in Fig. 4 *C*), this increase is smaller than the C_m increase for an NS-type repeat (compare $N(NS)_5C$ and $N(NS)_6C$ in Fig. 4 *A*). This suggests that addition of an NS-type repeat is more stabilizing than the addition of an HD-type repeat.

The suggestion that NS-type repeats are more stabilizing than HD-type repeats appears to be at odds with the observation that $(HD)_6C$ is more stable than $(NS)_6C$ (Fig. 3). One possible explanation for this apparent discrepancy is that stability differences between NS- and HD-type repeats are unevenly distributed between intrinsic folding and interfacial interaction energies.

Stabilities of TALEs containing mixtures of NS and HD TALE repeats

Naturally occurring TALE proteins contain "mixed RVDs," meaning that adjacent repeats have different RVD sequences. Fig. S6 shows urea-induced unfolding transitions for constructs containing both the NS and HD RVDs, $(HD)_2(NS)_1(HD)_2C$ and $(NS)_1(HD)_5C$. Both mixed RVD constructs have cooperative urea-induced unfolding transitions (Fig. S6), with *m*-values similar to constructs composed solely of one type of repeat. These observations suggest that the size of the cooperative unit is similar for mixed and unmixed constructs.

Global Ising analysis of NS and HD TALE repeat unfolding transitions

To dissect contributions of intrinsic and interfacial stabilities to repeat-protein folding, 1D Ising models were fitted to urea-induced unfolding transitions (19,20,22). These models represent unfolding at the level of individual repeats, and account for all combinations of folded and unfolded repeats. The free energy of each of these 2^n configurations (where *n* is the number of repeats) is modeled to the sum of the intrinsic energies of each folded repeat and the interfacial interaction energies of pairs of adjacent folded repeats. Because the caps differ in sequence, intrinsic energies are modeled as different from those of the internal, sequence-identical repeats.

Figs. 4 and S6 show a global fit of the Ising model to unfolding transitions for NS and HD constructs of different lengths and capping structures in which all NS-, HD-, and mixed RVD constructs are fitted simultaneously. Although each unfolding transition has separate baseline parameters, the entire family of transitions share 11 globally fitted thermodynamic parameters (Table 1). To estimate parameter uncertainties, 2000 iterations of bootstrapping were performed, and 95% confidence intervals were calculated (26).

All repeats have unfavorable intrinsic folding free energies and favorable interfacial free energies, consistent with previous Ising analyses of ankyrin repeat and some TPR proteins (19,20,22,23). Intrinsic folding of individual NS repeats is more unfavorable than repeats with the HD RVD. In contrast, adjacent NS repeats are more strongly coupled than adjacent HD repeats. For interfaces between

repeats with different RVD sequences, coupling is the same as for the first repeat type in the pair. That is, the identity of the RVD determines the coupling energy to the next repeat (but not the previous repeat).

DISCUSSION

Analysis of unfolding transitions of TALE constructs of different length and RVD sequence using a nearest-neighbor Ising model provides information on coupling energies, local folding free energies, and RVD sequence-stability correlation. We find that although changes in the residues responsible for conferring DNA binding specificity have moderate effects on the constructs studied here, these sequence changes have large effects on the distribution of stability within and between repeats, and on the cooperativity of TALE repeat arrays. These results suggest that TALE proteins used for genome editing have different local as well as global stabilities based on the RVDs chosen for DNA sequence recognition.

The identity of the RVD affects stability and cooperativity

Ising analysis of the TALE constructs in this study reveals intrinsically unstable repeat units coupled by stabilizing interfaces, consistent with studies from other linear repeat proteins. This partitioning results in an overall cooperative folding transition, and gives rise to an increase in native-state stability with repeat number. However, the magnitude of the partitioning varies depending on RVD-type. For NS RVDs the intrinsic folding free energy is +5.9 kcal/mol, and the interfacial free energy is −7.8 kcal/mol. For HD RVDs the intrinsic folding energy is +3.5 kcal/mol, and the interfacial energy is −5.0 kcal/mol. That is, individual folded NS-type repeats are more intrinsically unstable, but couple more strongly with their folded neighbors. One result of this

difference in partitioning is that NS-type repeats are more strongly coupled than HD-type repeats.

Because naturally occurring TALE proteins are composed of many different RVD sequences, "mixed interfaces" are formed between repeats with different RVDs. Our Ising analysis of mixed NS and HD TALE RVDs shows that for a pair of mixed RVDs, the interfacial energy is determined by the N-terminal repeat. That is, $\Delta G_{NS,i:HD,i+1} = \Delta G_{NS,i:NS,i+1}$ and $\Delta G_{HD,i:NS,i+1} = \Delta G_{HD,i:HD,i+1}$ (Table 1).

Weak coupling of the N-terminal cap

Although the conserved N-cap is required for DNA binding and transcription activation of the naturally occurring PthXo1 TALE array (8,13,31,33–35), we find that this cap only modestly enhances stability of the central repeats. Our Ising analysis is consistent with this observation, showing that the N-cap is intrinsically stable (−4.5 kcal/mol), but it is weakly coupled (−0.82 kcal/mol) to the central repeats. In one proposed mechanism for DNA binding, the N-cap nonspecifically binds DNA and facilitates diffusion (31,36). Weak coupling of the N-cap from the central repeats could uncouple nonspecific diffusive association from tight sequence-specific DNA binding of central repeats. In such a model, the N-cap acts to increase local concentration of TALEs on DNA while the central repeat domain can separately search for specific sequences.

TALE arrays significantly populate partly folded states

Proteins fold in a highly cooperative manner, rarifying populations of partially folded states. The Ising model allows us to determine the populations of partly folded states. These populations can be represented as a distribution of the local folding free energy of each repeat as a function of position (Fig. 5). We define folding free energy as the free energy of all states in which a given repeat is folded minus that of all
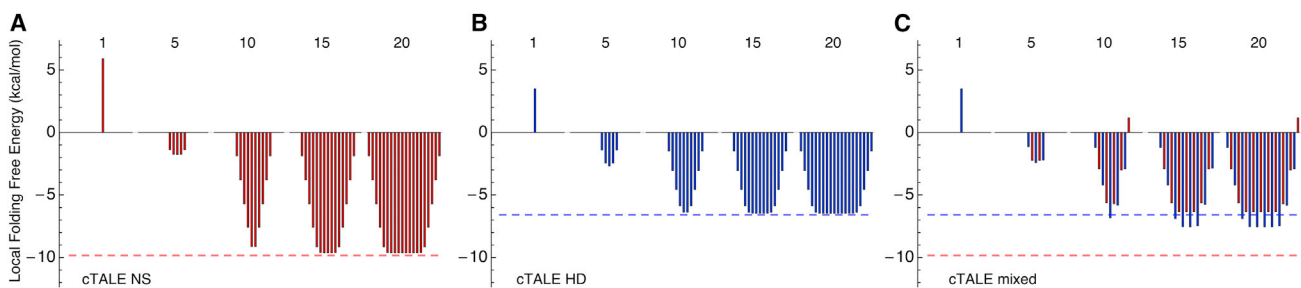

FIGURE 5 Distribution of local folding free energies for TALE repeat arrays as a function of length. At each position, probabilities of a repeat being folded (and unfolded) were calculated from the nearest-neighbor partition function using free energies from the Ising fit. Local folding free energies (ΔG°_{local}) were calculated from Eq. 4. For homopolymeric TALE arrays (A, NS-type repeats in *red*, B, HD-type repeats in *blue*), the ΔG°_{local} values of the central repeats decreases with repeat number until a plateau in local folding free energy is reached (*dashed lines*). This plateau in ΔG°_{local} is lower (i.e., greater stability) for NS repeats (A) than HD repeats (B). Constructs containing both NS and HD repeat types have a heterogeneous stability distribution, with local unfolding of the C-terminal NS repeats (C). Mixing NS (*red*) and HD (*blue*) repeats in an alternating fashion decreases ΔG°_{local} of HD repeats, but increases ΔG°_{local} of NS repeats. All constructs show fraying of end repeats, as would be expected from a cooperative nearest-neighbor model. To see this figure in color, go online.

states where the repeat is unfolded (see Supporting Materials and Methods). In addition to providing a picture of end-fraying, these ΔG_{local}° distributions provide a picture of accessibility of conformations that are unfolded (i.e., "broken") in the middle of the array.

For short (one to four repeat) homopolymeric TALE arrays (Fig. 5, A and B), ΔG_{local}° is positive, meaning it is more probable for repeats to be unfolded than folded. For homopolymer constructs with five or more repeats, ΔG_{local}° becomes negative. Central repeats have more negative ΔG_{local}° values than terminal repeats, consistent with end fraying. Somewhat surprisingly, the local folding free energies of internal repeats reach a plateau for long TALE arrays (15 and 20 repeats). The local folding free energies plateau at a lower value for NS-type repeats, that is, central NS-repeats are more stable than central HD-repeats (*dashed lines*, Fig. 5, A and B).

The values of local folding free energy plateaus are nearly equal to the sum of a single intrinsic energy and two interfacial energies (because two interfaces must be disrupted to fold an internal repeat). The close agreement between these two quantities indicates that the dominant partially unfolded state captured by the Ising model is one in which a single internal repeat unfolds, while the remainder of the TALE array remains folded. This kind of local "break" in the TALE array should disrupt the superhelix and allow direct DNA binding.

Because the intrinsic and interfacial energies for NS- and HD-type repeats are different, the plateau values for NS- and HD-type repeats are different (-9.7 kcal/mol for NS-type repeats and -6.6 kcal/mol for HD-type repeats, Fig. 5, A and B). Thus, the probability of a local unfolding reaction inside of the TALE array depends significantly (by a factor of $10^{3.1/RT}$) on RVD sequence.

Fig. 5 C shows the calculated ΔG_{local}° distribution for mixed constructs. Alternating NS- and HD-type repeats leads to significant heterogeneity in the ΔG_{Local}° distribution due to the difference in the intrinsic stabilities of NS- and HD-type repeats. Compared with unmixed arrays, mixing NS- and HD-type repeats increases the local folding free energy of NS-type repeats while decreasing the local folding free energy of HD-type repeats (Fig. 5 C). NS repeats have higher ΔG_{local}° values in the mixed repeat array because the N-terminal interface in the mixed system is less favorable. HD repeats have lower ΔG_{local}° values in the mixed repeat array because the N-terminal interface in the mixed system is more favorable.

Natural TALEs, such as PthXo1, contain a diverse set of RVD sequences. This extensive mixing of repeats may result in an increase in the population of partly folded states, either through end fraying, internal repeat unfolding, or interfacial fracture (see below and Fig. S7). The far UV CD spectrum of heteropolymeric PthXo1 has less α-helical signal compared with the far UV CD spectra of homopolymeric cTALEs (Fig. S4).

Regions of local instability may be important for facilitating binding to DNA.

TALEs access "fractured" states

A more limited type of local structural distortion is the disruption of a single interface between two folded repeats, which can be viewed as a break internal to the folded TALE array. Such states are not included in the Ising model, which assumes that adjacent folded repeats are automatically coupled through interfacial interaction. To capture these fractures, we modified the Ising partition function to include these states (Supporting Materials and Methods). The probability of fracture is calculated for arrays of 20 cTALE repeats (homopolymeric and mixed arrays, Fig. S7). Mixed cTALEs have the greatest fracture probability, whereas homopolymeric NS cTALEs have the lowest fracture probability (3.9×10^{-3} versus 1.9×10^{-5}, respectively). For comparison, the probability of fracture is calculated for consensus Ankyrin repeats (cANKS), another helical 34 residue linear repeat array that has been analyzed using the Ising formalism (19,22); arrays of cTALEs have fracture probabilities that are four to six orders of magnitude greater than arrays of cANKs.

We have described several different ways cTALEs break: end fraying, unfolding of internal repeats, and rupturing of interfaces. Calculated probabilities for these different types of breakage are compared in Fig. S7. States where terminal repeats are unfolded (end frayed) have the greatest probability. For cTALEs, states where internal repeats are unfolded (internally unfolded) and states where consecutive repeats are folded but uncoupled (interfacially fractured) are significantly populated. These types of structural distortion are energetically accessible to cTALEs and may provide access to a state that is competent for DNA binding, thus increasing the overall binding rate.

Length-dependent stability switch

Because of differences in the energetic partitioning for the NS and HD RVDs, the relative stabilities of these two arrays are predicted to switch as repeats are added. At low repeat number, constructs containing NS-type repeats are less stable than constructs containing the HD RVD. This can be seen in Fig. 3, where the midpoint of the $(NS)_6C$ transition is at lower urea concentration than that of $(HD)_6C$.

However, at high repeat number, the Ising model predicts that constructs composed of HD RVDs are less stable than constructs composed of NS RVDs. Although we have been unable to purify these longer constructs ($N(HD)_7$ and $N(HD)_8$ aggregate according to sedimentation velocity), we can use Ising parameters to estimate the free energy of the native state. (Fig. 6). At one repeat, the stability is equal to the intrinsic stability; thus, the "native state" of a single HD-repeat construct has a lower free energy than a single
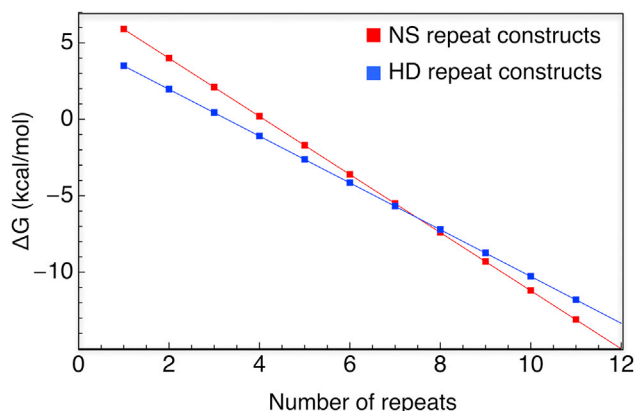
FIGURE 6 Differences in energetic partitioning for NS- and HD-RVDs results in a length dependent stability switch. Using free energies from the Ising fit, folding free energies for arrays with increasing repeat number were calculated. At low repeat number, the fully folded state of NS-type constructs are less stable than HD-type constructs. However, at high repeat number, NS-type constructs are more stable than HD-type constructs. The crossover point is between seven and eight repeats. To see this figure in color, go online.

NS-repeat construct. Upon the addition of subsequent repeats, native state free energy decreases linearly, with a slope equal to the sum of the intrinsic and interfacial energies. Because this sum is larger (more negative) for NS repeats, the two free energy lines intersect between seven and eight repeats.

CONCLUSIONS

Through manipulation of capping sequences, we have designed a consensus TALE array that permits both length and RVD sequence variation, and we have found conditions where folded constructs remain monomeric, even with only one terminal cap. Together, our set of constructs satisfies the criteria for 1D Ising analysis, permitting high-precision determination of intrinsic stabilities and nearest-neighbor coupling energies. All constructs show moderate cooperativities (favorable coupling energies, unfavorable intrinsic stabilities). Importantly, these parameters are strongly dependent on RVD sequence. Although we have only looked at NS- and HD-RVDS, it is clear from our studies that RVD sequence identity influences not only global stability, but also intrinsic stability, coupling energy, cooperativity, and accessibility of partly folded states. Together, these factors have implications for genome editing: depending on the sequence of the genomic target, the thermodynamic profile of the cognate TALE may affect the ability to activate target sites. Taking these factors into account in genome editing experiments could improve activity.

SUPPORTING MATERIAL

Supporting Materials and Methods and seven figures are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(16)30940-7.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

REFERENCES

1. Kay, S., S. Hahn, …, U. Bonas. 2007. A bacterial effector acts as a plant transcription factor and induces a cell size regulator. *Science.* 318:648–651.

2. Römer, P., S. Hahn, …, T. Lahaye. 2007. Plant pathogen recognition mediated by promoter activation of the pepper Bs3 resistance gene. *Science.* 318:645–648.

3. Boch, J., and U. Bonas. 2010. Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annu. Rev. Phytopathol.* 48:419–436.

4. Boch, J., H. Scholze, …, U. Bonas. 2009. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science.* 326:1509–1512.

5. Moscou, M. J., and A. J. Bogdanove. 2009. A simple cipher governs DNA recognition by TAL effectors. *Science.* 326:1501.

6. Miller, J. C., L. Zhang, …, E. J. Rebar. 2015. Improved specificity of TALE-based genome editing using an expanded RVD repertoire. *Nat. Methods.* 12:465–471.

7. Li, T., S. Huang, …, B. Yang. 2011. TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Res.* 39:359–372.

8. Christian, M., T. Cermak, …, D. F. Voytas. 2010. Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics.* 186:757–761.

9. Cong, L., R. Zhou, …, F. Zhang. 2012. Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nat. Commun.* 3:968.

10. Geissler, R., H. Scholze, …, J. Boch. 2011. Transcriptional activators of human genes with programmable DNA-specificity. *PLoS One.* 6:e19509.

11. Li, Y., R. Moore, …, L. Bleris. 2012. Transcription activator-like effector hybrids for conditional control and rewiring of chromosomal transgene expression. *Sci. Rep.* 2:897.

12. Mahfouz, M. M., L. Li, …, J.-K. Zhu. 2012. Targeted transcriptional repression using a chimeric TALE-SRDX repressor protein. *Plant Mol. Biol.* 78:311–321.

13. Zhang, F., L. Cong, …, P. Arlotta. 2011. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat. Biotechnol.* 29:149–153.

14. Maeder, M. L., J. F. Angstman, …, J. K. Joung. 2013. Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nat. Biotechnol.* 31:1137–1142.

15. Miyanari, Y., C. Ziegler-Birling, and M.-E. Torres-Padilla. 2013. Live visualization of chromatin dynamics with fluorescent TALEs. *Nat. Struct. Mol. Biol.* 20:1321–1324.

16. Ma, H., P. Reyes-Gutierrez, and T. Pederson. 2013. Visualization of repetitive DNA sequences in human chromosomes with transcription activator-like effectors. *Proc. Natl. Acad. Sci. USA.* 110:21048–21053.

17. Kim, H., and J.-S. Kim. 2014. A guide to genome engineering with programmable nucleases. *Nat. Rev. Genet.* 15:321–334.

18. Deng, D., C. Yan, …, N. Yan. 2012. Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science.* 335:720–723.

19. Aksel, T., A. Majumdar, and D. Barrick. 2011. The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. *Structure.* 19:349–360.

20. Kajander, T., A. L. Cortajarena, …, L. Regan. 2005. A new folding paradigm for repeat proteins. *J. Am. Chem. Soc.* 127:10188–10190.

21. Mello, C. C., and D. Barrick. 2004. An experimentally determined protein folding energy landscape. *Proc. Natl. Acad. Sci. USA.* 101:14102–14107.

22. Wetzel, S. K., G. Settanni, …, A. Plückthun. 2008. Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *J. Mol. Biol.* 376:241–257.

23. Marold, J. D., J. M. Kavran, …, D. Barrick. 2015. A naturally occurring repeat protein with high internal sequence identity defines a new class of TPR-like proteins. *Structure.* 23:2055–2065.

24. Cermak, T., E. L. Doyle, …, D. F. Voytas. 2011. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.* 39:e82.

25. Poland, D., and H. A. Scheraga. 1970. Theory of Helix-Coil Transitions in Biopolymers: Statistical Mechanical Theory of Order-Disorder Transitions in Biological Macromolecules. Academic Press, San Diego, CA.

26. Aksel, T., and D. Barrick. 2009. Analysis of repeat-protein folding using nearest-neighbor statistical mechanical models. *Methods Enzymol.* 455:95–125.

27. Finn, R. D., J. Clements, and S. R. Eddy. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39:gkr367.

28. Wheeler, T. J., J. Clements, and R. D. Finn. 2014. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics.* 15:7.

29. Mosavi, L. K., and Z.-Y. Peng. 2003. Structure-based substitutions for increased solubility of a designed protein. *Protein Eng.* 16:739–745.

30. Tripp, K. W., and D. Barrick. 2007. Enhancing the stability and folding rate of a repeat protein through the addition of consensus repeats. *J. Mol. Biol.* 365:1187–1200.

31. Gao, H., X. Wu, …, Z. Han. 2012. Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell Res.* 22:1716–1720.

32. Mak, A. N.-S., P. Bradley, …, B. L. Stoddard. 2012. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science.* 335:716–719.

33. Miller, J. C., S. Tan, …, E. J. Rebar. 2011. A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.* 29:143–148.

34. Mussolino, C., R. Morbitzer, …, T. Cathomen. 2011. A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res.* 39:9283–9293.

35. Sun, N., J. Liang, …, H. Zhao. 2012. Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. *Mol. Biosyst.* 8:1255–1263.

36. Meckler, J. F., M. S. Bhakta, …, E. P. Baldwin. 2013. Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic Acids Res.* 41:4118–4128.

Supplemental Information

Broken TALEs: Transcription Activator-like Effectors Populate Partly Folded States
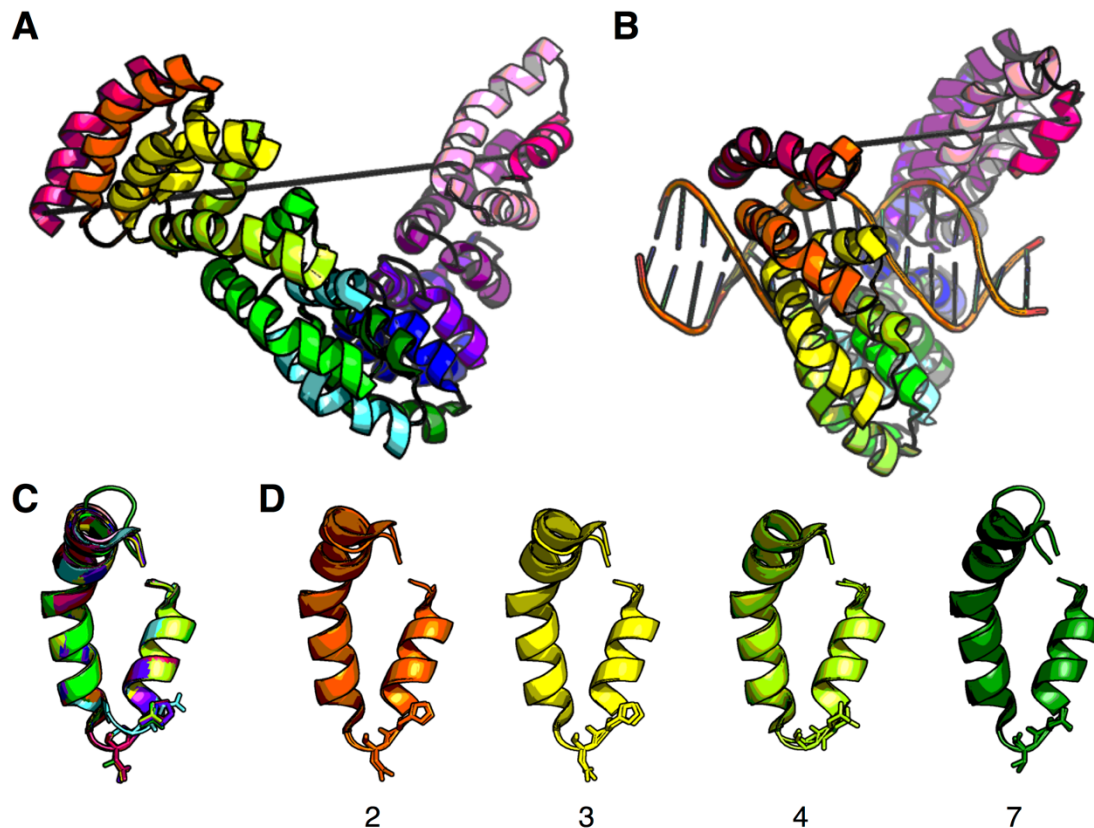
Kathryn Geiger-Schuller and Doug Barrick

Figure S1. The TALE conformational change upon DNA binding is mediated by small changes propagated through many repeats. (A) DNA-free structure of dHax3 (colored by repeat, PDB:3V6P). The distance between C_α of residue 304 to C_α of residue 666 is 74 Å. (B) DNA-bound structure of dHax3 (PDB:3V6T). The distance between C_α of residue 304 to C_α of residue 666 is 50 Å. (C) Alignment of C_αs of all 11 repeats in the DNA-bound structure. RMSDs range from 0.3 to 0.6 $Å^2$. (D) Alignment of C_αs of repeats 2, 3, 4, and 7 in the DNA-free and DNA-bound structure. RMSDs range from 0.3 to 0.6 $Å^2$.
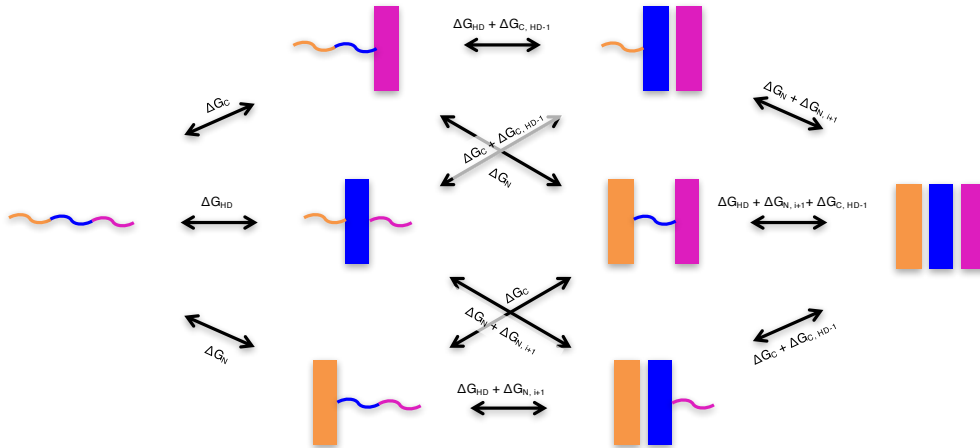
Figure S2. A 1-D Ising Model for a cTALE with three repeats. In this example, a construct with the one N-cap, one HD repeat, and one C-cap is shown. The partition function q is used to derive the dependence of the fraction folded repeats (f_{folded}) on urea concentration (See equations 1-4). The folding of each repeat results in an intrinsic free energy change, along with free energy changes from formation of one or two interfaces (for terminal repeat unfolding and internal repeat folding respectively).
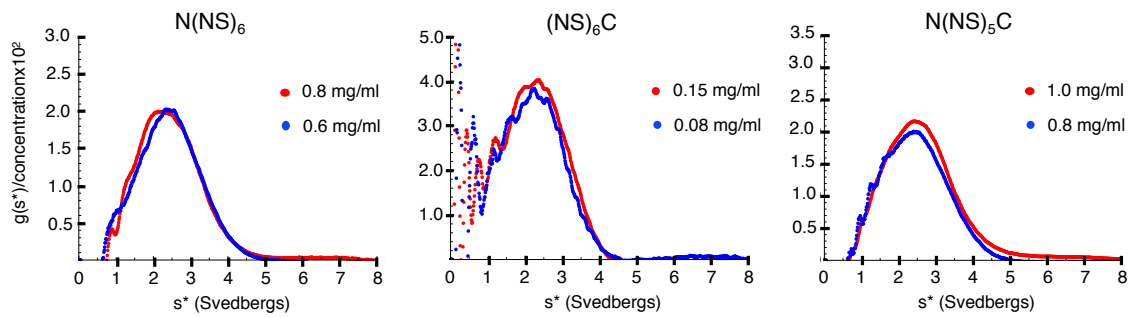
Figure S3. Sedimentation Velocity g(s*) plots for capped consensus TALE repeats. Sedimentation velocity experiments were performed with N(NS)$_6$, (NS)$_6$C, and N(NS)$_5$C, as described in Supporting Materials and Methods. The g(s*) distributions are consistent with monomers. Conditions: 300 mM NaCl, 10 mM Tris HCl pH 7.4, 5% glycerol, 25°C.
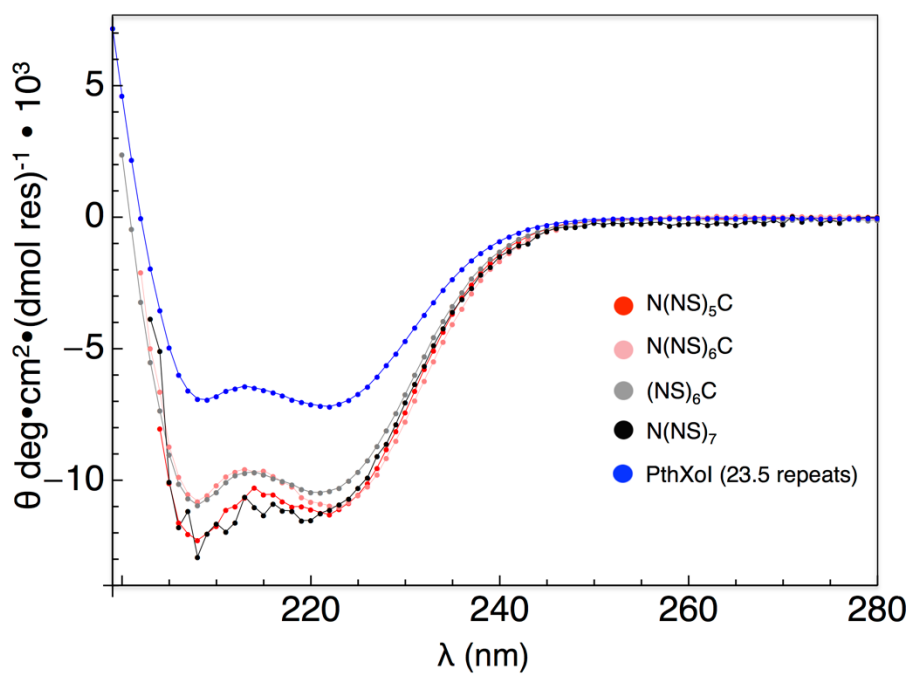
Figure S4. Consensus and natural TALEs and have α-helical secondary structure. FarUV CD of cTALEs and the repeat domain of PthXo1, a naturally-occurring TALE, show similar shape. PthXo1 has less helical structure than the cTALEs.

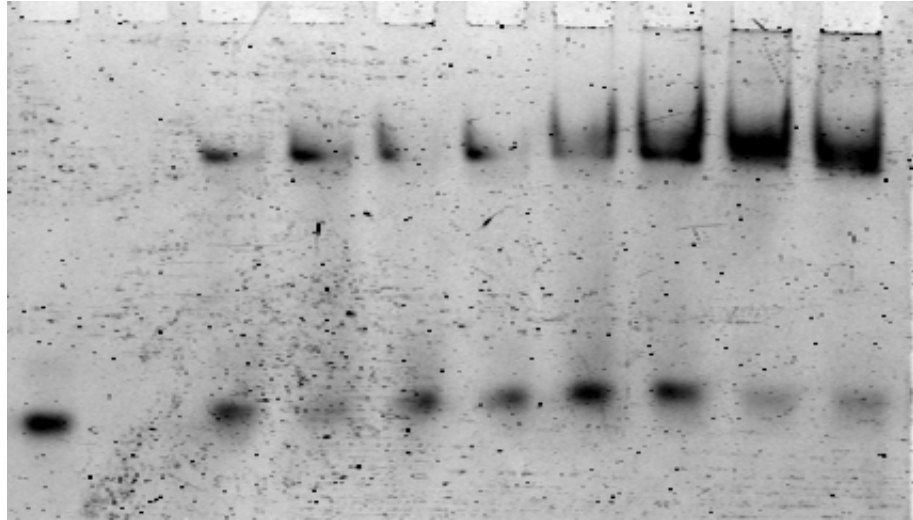| N(NS)$_6$C (nM)	0	0	10	30	60	100	300	500	800	1000
fam-dsA$_{15}$/T$_{15}$	+	-	+	+	+	+	+	+	+	+

Figure S5. Consensus TALEs bind double stranded DNA. An electrophoretic mobility shift assay (EMSA) shows N(NS)$_6$C binding 10 nM fam-dsA$_{15}$/T$_{15}$ (see Supporting Materials and Methods). Conditions: 150 mM KCl, 0.1 mM DTT, 10 mM Tris, pH 7.4, 25% sucrose.
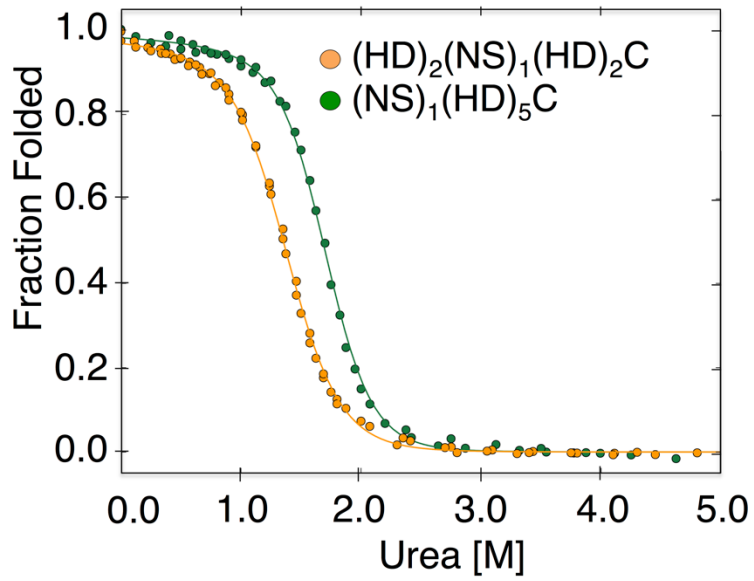
Figure S6. "Mixed RVD" constructs have single cooperative unfolding transition. Urea-induced unfolding transitions for constructs containing both NS- and HD-type repeats. Transitions are cooperative and slopes of $(HD)_2(NS)_1(HD)_2C$ and $(NS)_1(HD)_5C$ are similar to $(NS)_6C$, suggesting cooperative unfolding units of similar size.

	$p_{\text{fully folded}}$	$p_{\text{end frayed}}$	$p_{\text{ internally unfolded}}$	$p_{\text{ interfacially fractured}}$
cTALE (NS)	0.92	0.08	8.3×10^{-7}	1.9×10^{-5}
cTALE(HD)	0.85	0.15	1.7×10^{-4}	3.9×10^{-3}
cTALE(mix)	0.11	0.89	1.1×10^{-4}	3.1×10^{-3}
cANK	>0.999	<0.001	1.5×10^{-14}	8.4×10^{-9}

fully folded   end frayed   internally unfolded   interfacially fractured

Figure S7. Calculated probabilities for fully folded and broken TALEs.
Probabilities of fully folded, end frayed, internally unfolded, and interfacially fractured populations were calculated using free energies from Ising analysis of cTALEs and, for comparison, consensus ankyrin repeat proteins (cANK)(19). Calculations were performed on arrays of 20 repeats. Microstates containing unfolded repeats in the regions 1-5 or 16-20 were included in the calculation of $p_{\text{end-frayed}}$. Microstates containing unfolding in repeats 6-15 were included in the calculation of $p_{\text{internally unfolded}}$. To calculate $p_{\text{interfacially fractured}}$, a separate Ising-like model was generated that includes microstates with structural deformation between two consecutive uncoupled but folded repeats (Supporting Materials and Methods). There is a significant population of broken states for cTALEs, in contrast to the very low probabilities of broken states calculated for cANKs.

Supporting Materials and Methods

Sedimentation Velocity

Analytical ultracentrifugation sedimentation velocity experiments were performed using a Beckman XL-1 analytical ultracentrifuge as previously described(22). Proteins were dialyzed in reference buffer for 24 hours prior to centrifugation. Concentrations ranged from 2 to 30 µM. AUC data were analyzed with SEDANAL. DCDT analysis was performed on scans spanning a one hour interval, generating s* value distributions (g(s*) versus s*)(23).

Electrophoretic Mobility Shift Assay (EMSA)

Flourescein (Fam)-labeled A_{15} single stranded DNA was annealed with unlabeled T_{15} single stranded DNA to prepare fam-dsA_{15}/T_{15}. Reactions with increasing concentrations of cTALEs were incubated at room temperature for 20 minutes. Samples were then loaded onto fresh 6% non-denaturing 0.5X TBE gels and were electrophoresed at 100 volts in the cold room for 60 minutes. Gels were imaged on the JHU Integrated Imaging Center Typhoon 9410 Variable Mode Imager and analyzed with ImageJ.

Calculating local folding free energy from intrinsic and interfacial stabilities

Local folding free energies (ΔG°_{local}) describe the free energy difference between all states where a particular repeat, *i*, is folded, and all states where the ith repeat is unfolded (Equation S1). For example, to calculate the ΔG°_{local} of the third repeat in a three repeat array, partition functions for each construct are created as illustrated in Figure S5. Next, q_3 sums all statistical weights of all conformations where the third repeat is folded. Placing zeros in the third matrix is equivalent to not counting statistical weights when the third repeat is unfolded (Equation S2). Dividing q_3 by the partition function, q (Equation S3), gives the probability that the third repeat is folded (Equation S4). The probability that the third repeat is unfolded is simply one subtracted by the probability of the third repeat being folded. The ΔG°_{Local} of a repeat is the log of the ratio of these probabilities (Equation S5).

$$Local\ Stability = -RT \ln \left(\frac{\text{▐▌} + \text{〜▌} + \text{▐-▌} + \text{▐▌▌}}{\text{▐-} + \text{〜〜} + \text{▐〜} + \text{▐▌-}} \right) \tag{S1}$$

$$q_3 = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa\tau & \kappa \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa\tau & \kappa \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa\tau & \kappa \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{S2}$$

$$q = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa\tau & \kappa \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa\tau & \kappa \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa\tau & \kappa \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{S3}$$

$$p_{3,N} = \frac{q_3}{q} \tag{S4}$$

$$\Delta G^{\circ}_{Local} = -RT \ln \left(\frac{p_{3,N}}{p_{3,D}} \right) \tag{S5}$$

An extended Ising model to include fractured states

      The 1-D Ising model we used to describe folding can be extended to include states where consecutive repeats are folded but uncoupled. The physical interpretation of such states is that a structural deformation occurs leaving repeats folded but preventing them from forming favorable interactions required for coupling. Such states should be accessible by opening a turn between helices, requiring changes to just a few backbone dihedral angles. Generation of partition functions for the folding Ising model is described in Equations 1-4 in the Materials and Methods. Equation 2 can be modified to include fractured states by replacing $\kappa\tau$ with $\kappa\tau + \kappa$ in the first column and row of the 2x2 matrices for each repeat.

$$\theta = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_N \tau_N + \kappa_N & \kappa_N \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_R \tau_R + \kappa_R & \kappa_R \\ 1 & 1 \end{bmatrix}^{n-2} \begin{bmatrix} \kappa_C \tau_C + \kappa_C & \kappa_C \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad \text{(S6)}$$

      If the i and $i + 1$ repeats can be folded or unfolded, there are four possible states (FF, FU, UF, and UU). Each of the four elements in the 2x2 matrices represent one of these states.   Row one, column one in each matrix represents states where the i and $i + 1$ repeats are folded (FF). In the folding Ising model, if i and $i + 1$ are folded, the statistical weight is $\kappa\tau$ because the interfacial energy is automatically associated with the folding of the ith repeat (through τ). To allow for states where there is no coupling between consecutive repeats, we simply add a term κ to the position where both repeats are folded. This is saying that when the i and $i + 1$ repeats are folded (FF), either there is a coupling ($\kappa\tau$) or there is not (κ). Because these two options are mutually exclusive, their contributions to statistical weight sum.

      Urea induced unfolding transitions fitted with this extended model returned best-fit values within 0.03% of the best-fit values in Table 1, with the exception of the interfacial energy of the N-cap and next repeat ($\Delta G_{N, N+1}$). Because the interface between the N-cap and the adjacent repeat is rather weak, the sum $\kappa\tau$ $+\kappa$ in the modified Ising model is significantly larger than the Ising value $\kappa\tau$. As a result, the best-fit value for $\Delta G_{N, N+1}$ is -0.69 kcal/mol with the fracture Ising model as compared to -0.85 kcal/mol with the folding Ising model.

Supporting References

22.   Marold, J.D., J.M. Kavran, G.D. Bowman, and D. Barrick. 2015. A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins. Struct. Lond. Engl. 1993. .

23.   Stafford, W.F. 1992. Boundary analysis in sedimentation transport experiments: a procedure for obtaining sedimentation coefficient distributions using the time derivative of the concentration profile. Anal. Biochem. 203: 295–301.