PCCP

Competition of individual domain folding with inter-domain interaction in WW domain engineered repeat proteins

Journal:	*Physical Chemistry Chemical Physics*
Manuscript ID	CP-ART-12-2018-007775.R2
Article Type:	Paper
Date Submitted by the Author:	05-Oct-2019
Complete List of Authors:	Dave, Kapil; IISER Mohali, Chemical Sciences Gasic, Andrei; University of Houston, Physics Cheung, Margaret S.; University of Houston, Physics Gruebele, Martin; University of Illinois at Urbana Champaign, Chemistry

Submitted to: PhysChemChemPhys

Competition of individual domain folding with inter-domain interaction in WW domain engineered repeat proteins

Kapil Dave,[a,†] Andrei G. Gasic,[b,c,†] Margaret S. Cheung[b,c,★] and M. Gruebele[a,d,★]

[a]Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Champaign, IL, 61801, USA.
[b]University of Houston, Department of Physics, Houston, Texas, 77204, USA.
[c]Center for Theoretical Biological Physics, Rice University, 77005, USA.
[d]Department of Physics and Department of Chemistry, University of Illinois at Urbana-Champaign, Champaign, IL 61801, USA.

[†]These authors contributed equally to this work.

[★]Corresponding author emails: mgruebel@illinois.edu; mscheung@uh.edu

Electronic Supplementary Information (ESI) available. The PDF file contains 8 sections with additional figures, formulas and parameter lists for the manuscript. See DOI: 10.1039/XXX.

Abstract Engineered repeat proteins have proven to be a fertile ground for studying the competition between folding, misfolding and transient aggregation of tethered protein domains. We examine the interplay between folding and inter-domain interactions of engineered FiP35 WW domain repeat proteins with n = 1 through 5 repeats. We characterize protein expression, thermal and guanidium melts, as well as laser T-jump kinetics. All experimental data is fitted by a global fitting model with two states per domain (U, N), plus a third state M to account for non-native states due to domain interactions present in all but the monomer. A detailed structural model is provided by coarse-grained simulated annealing using the AWSEM Hamiltonian. Tethered FiP35 WW domains with n=2 and 3 domains are just slightly less stable than the monomer. The n=4 oligomer is yet less stable, its expression yield is much lower than the monomer's, and depends on the purification tag used. The n=5 plasmid did not express at all, indicating sudden onset of aggregation past n=4. Thus, tethered FiP35 has a critical nucleus size for inter-domain aggregation of $n \approx 4$. According to our simulations, misfolded structures become increasingly prevalent as one proceeds from monomer to pentamer, with extended inter-domain beta sheets appearing first, then multi-sheet 'intramolecular amyloid' structures, and finally novel motifs containing alpha helices. We discuss the implications of our results for oligomeric aggregate formation and structure, transient aggregation of proteins whilst folding, as well as for protein evolution that starts with repeat proteins.

1. Introduction

An important question in protein dynamics is how proteins manage to fold in the presence of many other biomolecules with which they could interact instead inside the cell.[1] For example, it has been discussed extensively how proteins can transiently aggregate during their folding process, thus mimicking the existence of monomeric folding intermediates.[2]

Repeat proteins are particularly interesting subjects for studying the interplay between folding and aggregation.[3] The proximity of tethered domains with identical or near-identical folds enhances protein-protein interactions.[4,5] It was shown by Borgia *et al.* for immunoglobulin-like oligomeric repeats that identical neighbors transiently misfold more readily than neighbors of lower sequence identity.[6,7] Thus evolutionary pressure reduces sequence similarity between adjacent repeat domains, and many natural repeat proteins contain folds that do not interact too strongly. Such sequences can also be engineered: the energy landscape of some ankyrin repeat proteins, especially of consensus[8] sequences, enables parallel folding of the domains.[9] We showed in our study of identical tethered repeats of protein U1A that increasing the number of repeat units increases transient aggregation relative to folding, until a well-defined nucleus for oligomeric aggregate formation is reached.[10,11] Javadi and Main observed similar frustration when increasing repeat number of the helical tetratricopeptide from 2 to 10,[12] and the energy landscape model can explain such frustration.[13,14] There is also structural information about misfolded states of repeat proteins. For example, immunoglobulin-like repeats[15] can form long-lived domain-swapped structure, and transient 'intramolecular amyloids' as previously postulated for lambda repressor fragment experimentally[16] and for other proteins computationally.[17] The computational study by Tian and Best[18] rationalizes domains-swapping as a late-stage process during (mis)folding.

Although there may be evolutionary pressure against repeat proteins due to inter-domain aggregation, in addition to titin,[6] IκB,[5] and other ankyrins,[9,19] repeats occur for many other natural proteins.[20] For example, tandem repeats of WW domains improve cell regulation. Schueler-Furman and co-workers published a detailed overview on how WW tandem repeat modules facilitate fine-tuning of regulation inside cells.[21] Possible ways include: 1) Repeat domains contribute to an overall increase in binding affinity. 2) One domain assists the binding of another (chaperone effect) 3) Adjacent WW domains can change the dynamics and stability of the neighboring domain. The vital role in cellular regulation of the family of tandem repeat WW domains provides additional motivation to investigate such a system in mechanistic detail.

Here we study the competition between folding and transient (or permanent) aggregation of

engineered WW domain oligomers, with n = 1-5 domains tethered by short serine-glycine sequences. We chose "GS linkers" because they have been well characterized, are highly soluble, and allow monomers to interact.[22] Experimentally, we study how increasing numbers of repeats affect expression levels, secondary structure, and transient misfolding. We find that at n = 4, misfolded oligomers strongly compete with native folding, and that n = 5 represents a hard limit for expression under our experimental conditions. The results are well-modeled by a global fit to a domain-state model. We complement the experiments and domain-state model with coarse-grained simulated annealing simulations to obtain structural information about the misfolded oligomers. A variety of interesting structures emerges, from individual misfolded domains, to chimeric misfolds (where two proteins intermingle), to entirely new beta-sheet structures, and finally even to alpha-helical structures that bear no resemblance to the original domains making up the oligomer.

2. Methods

2.1 Protein sample preparation We studied monomer to pentamer repeat sequences of FiP35 WW domain, each domain unit being separated by a 10-residue glycine-serine flexible linker. The leader (after cutting the hexa-histidine tag), monomer and linker sequences are

HM + KLPPGWEKRMSRDGRVYYFNHITNASQFERPSG + **GGSGGSGGSG**

Table S1 in the Supplementary Information section 1 shows the full sequences. The dimer through tetramer were cloned into pDream (GenScript) and expressed in BL21 (DE3)-RIPL *E. coli* (Agilent) cells. The constructs were purified on a Ni-NTA column using the His-tag via the same protocol previously used for His-tagged U1A oligomers.[11]

For comparison, the monomer and tetramer were also expressed as fusion proteins with a Glutathione-S-transferase (GST) tag and a thrombin cleavage site for purification as described previously.[23] A non-dimerizing GST tag (Genscript) was used, although residual interactions may persist. The fusion protein was captured and purified from the cell extract on an immobilized glutathione resin according to manufacturer's guidelines (GenScript): The protein was eluted by 10 mM glutathione in 50 mM Tris-HCl pH 8.0 and followed by dialysis in 10 mM sodium phosphate buffer. The purification tag was cleaved by overnight incubation with biotinylated thrombin (EMD Millipore). Thrombin was removed by incubation with streptavidin-agarose resin (EMD Millipore) according to manufacturer's protocol. The monomer was purified from cleaved GST via an ultrafiltration cell with 10 kDa cutoff membrane (Millipore). Due to comparable size

of GST and QFiP35, their separation was performed by passing the cleaved protein solution through a gravity column with immobilized glutathione resin.

The presence of a single tryptophan on the first β-strand (loop or hairpin 1) of each WW domain enabled monitoring of folding via fluorescence. The pentamer could not be expressed in sufficient yield for experiments.

2.2 Temperature unfolding thermodynamics Fluorescence spectroscopy was carried out using a JASCO fluorescence spectrophotometer FP-8300 equipped with programmable temperature control and excitation and emission slit widths at 5 nm wavelength resolution. Temperature denaturation of all constructs was measured by exciting tryptophan fluorescence at 280 nm and measuring the fluorescence wavelength shift. The tryptophan fluorescence peak red-shifts as the sidechain is exposed to a more polar solvent environment. We analyze wavelength shift because it is less dependent on quantum yield or concentration than fluorescence intensity. For each fluorescence emission spectrum, the average wavelength $\langle\lambda\rangle$ was calculated by equation (1) where I is intensity and λ wavelength:[24]

$$\langle\lambda\rangle = (\sum_j \lambda_j I_j) / (\sum_j I_j). \quad (1)$$

The average wavelength tracks the fluorescence peak wavelength as long as the same wavelength range is used consistently, but provides higher signal-to-noise ratio because it utilizes the full fluorescence spectrum, rather than just fitting a few points near the peak of the spectrum. The same wavelength range of 290 to 450 nm was used in all calculations to obtain consistent results. Temperature cycling of 40 μM protein solution from 20 °C to 90 °C and back, with or without GuHCl, produced very similar results as 1 μM solutions, recovering $\langle\lambda\rangle$ within *ca*. 3 nm. Thus there is inter-domain aggregation at 90 °C, but no concentration-dependent intermolecular aggregation up to 40 μM.

Circular dichroism was measured using a JASCO spectrophotometer with Peltier temperature control (JASCO Inc, Easton MD). All spectra were recorded from 250 – 200 nm at a scan rate of 50 nm/min with 1 nm resolution and are an average of 5 accumulations. We used a 1 mm path length quartz cuvette and, unless otherwise noted, at a protein concentration of 10 μM.

All thermal denaturation signals $S(T)$ in absence of denaturant were fitted to a two-state model for temperature denaturation

$$S(T) = S_U + S_F e^{-\Delta G(T)/RT} / (1 + e^{-\Delta G(T)/RT}) \quad (2a)$$

$$\Delta G(T) = g_T(T - T_m) \quad (2b)$$

to obtain the denaturation midpoints with respect to temperature (T_m). Different values of g_T and T_m are obtained at different denaturant concentrations, and this is reflected in the global fit described in section 2.5.

2.3 Temperature jump kinetics Laser temperature jumps were carried out using a Surelite Q-switched Nd:YAG laser (Continuum Inc., Santa Clara, CA), with details of the instrument mentioned elsewhere [25,26].The jump size was 5-6 °C. The exact size of the jump was calibrated by comparing the fluorescence decays *f* of tryptophan (300 μM solution) after the jump with the corresponding decay at an equilibrium temperature several degrees higher. Fluorescence decays were excited at 280 nm by a tripled, mode-locked Ti:sapphire laser every 12.5 ns for a total of 1 ms. The temperature jump was set to occur 153.75 μs after the oscilloscope was triggered to start data collection. The sampling frequency was 10 Giga-samples per second. Thus, each fluorescence decay was sampled at 100 picosecond intervals, or 125 times before the next decay was excited. The signal was usually 50-60 mV. Sample concentration was 40 μM for all of the proteins. At this concentration the thermal denaturation was reversible.

2.4 Kinetics analysis Kinetics data were analyzed using MATLAB (Mathworks Inc., Natick, MA) and IGOR Pro (Wavemetrics Inc., Lake Oswego, OR). A fluorescence decay $f(t)$ was collected every 12.5 ns. 100 of these were binned into intervals of 1.25 μs. Thus, the protein kinetics could be followed with 1.25 μs time resolution. We assumed a two-state kinetic model, which fitted the data within uncertainty. The decays $f(t)$ were fitted to a linear combination of the decay f_1 averaged between 153.75 and 28.75 μs before the T-jump, and the decay f_2 averaged over the final 125 μs of data collection, where the protein had equilibrated: $f(t) = a_1(t) f_1 + a_2(t) f_2$. The relative lifetime shift as a function of time is $\chi(t) = a_1(t)/[a_1(t) + a_2(t)]$, irrespective of how a_1 and a_2 are normalized. The $\chi(t)$ traces were fitted using the domain-state model described next.

2.5 Global fitting model with a misfolded state representing domain interaction Unlike U1A, which is prone to aggregation at > 1 μM protein concentration,[11] WW domain monomer MFiP35 is a fast folder that can be studied at >100 μM concentration. In addition to its native state N and unfolded state U, WW domain has only very short-lived (μs) on-pathway intermediates (one or the other hairpin folded).[27] Therefore domain interactions are weak, and we hypothesize that they can be accounted for by an additive free energy model with an extra state "M" added to represent non-native structure caused by domain interactions.

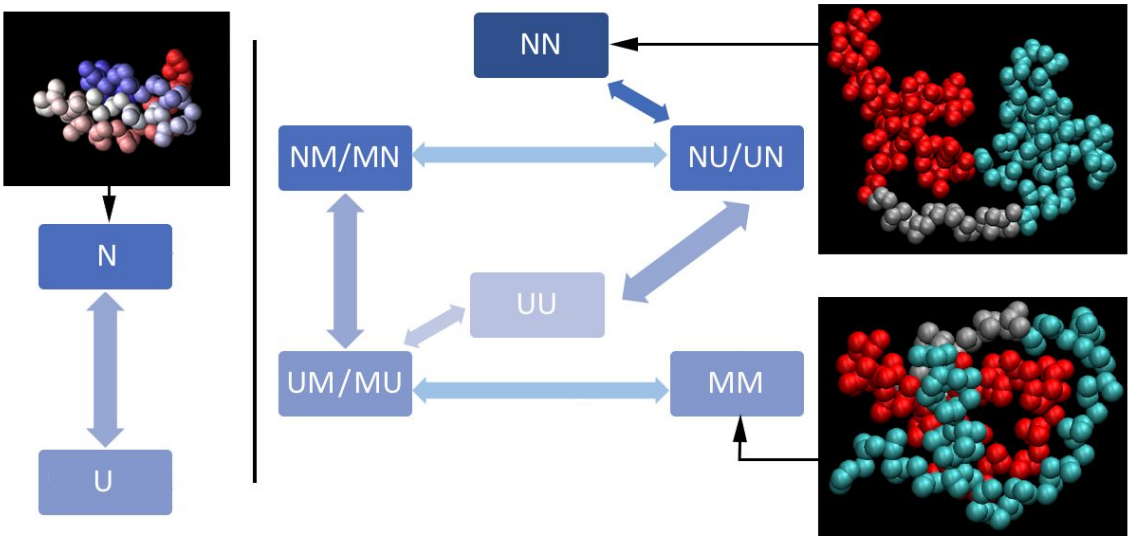

Fig. 1: Left: native structure and fitting model for the monomer. Middle and right: global fitting model and structures for the dimer. The allowed interconversions between the various species are shown using arrows. The kinetic model only flips one domain at a time, i.e. the model assumes that there are no simultaneous double transitions. Representative structures of the NN and MM states are from the coarse-grained simulations.

Our goal was to build a minimalist kinetic network model that can fit all measured equilibrium (temperature+denaturant) melts and kinetics data for all *n*-mers simultaneously. Each *n*-mer is modeled at the level of states for each individual domain. For example, the tetramer QFiP35 is represented as "XXXX". At a minimum, each domain "X" can attain one of the two forms: N (native) and U (unfolded). For all proteins *except* the monomer (Figure 1 left), when domains interact a third state M accounts for misfolded states due to domain interaction. The model does not explicitly treat which domains are interacting to put a domain in state M. Inter-domain interaction is judged by how well a two-state model (only states N and U for each domain) fits the data compared to the full model (extra state M to account for misfolding due to domain interactions except for the monomer).

The coupling between domains is thus implicit in pathways involving state M, such as those shown in Figure 1 (middle and right).; i.e., if all the reaction paths were possible, and the model explicitly had coupling between domains, the most probable paths would be the ones shown in Figure 1. Hence, our data fitting model uses population of the state M, instead of explicit interaction terms between N and U, to treat inter-domain interaction and the resulting non-native states. Our coarse-grained annealing simulations (sections 2.6-2.8 and 3.6-3.10), which explicitly contain inter-domain interactions, provide more insight into what a state such as "NMM" actually might look like.

The equilibrium population of each *n*-mer state as a function of temperature and denaturant concentration was calculated from the additive free energy function

$$\Delta G(T;\#M,\#N) = \#N\{g_{UN}(T - T_m) + m_{UN}[GuHCl]\} + \#M\{g_{UM}(T - T_m) + m_{UM}[GuHCl] + g^{(0)}_{UM}\} + \#U \cdot 0\ . \quad (3)$$

The terms in the free energy equation are defined as follows:

- #N, #M and #U equals the number of N, M, and U domains present
- T_m is the melting temperature of native domains in absence of denaturant or intermediates
- g_{UN} is the thermal free energy derivative (folded domain relative to unfolded domain)
- m_{UN} is the corresponding denaturant free energy derivative
- g_{UM} is the thermal free energy derivative (misfolded domain relative to unfolded domain)
- m_{UM} is the corresponding denaturant free energy derivative
- $g^{(0)}_{UM}$ is the free energy of M relative to U at T=T_m and no denaturant

Supplementary Information section 2 provides more details. The global model does not contain an explicit domain-domain interaction term, so the size-dependence of these interactions is included only in the term ~ $\#M$, which scales linearly with $\#M$.

The experimental fluorescence thermal melts did not show two separate step-wise transitions, only shifts of stability that could not be accounted for by only two states per domain. Hence our global model made the simplest assumption consistent with the data: the misfolded state's fluorescence signal (S_M) was assumed to be an average of the folded (S_F) and unfolded signals (S_U). This is a reasonable assumption: our experimentally observed signal is tryptophan fluorescence, and that residue is most buried in the folded state, least buried in the unfolded state, whereas the coarse-grained annealing simulations (see 2.6) show intermediate solvent exposure of the tryptophan in most of the modeled misfolded states. Picking other intermediate state signals S_M affects the precise populations of intermediate states, but not the general conclusions from our model.

Kinetics were similarly modeled by inclusion of two free energy barriers in the global fitting model. In order for the model to mimic the T-jump relaxation experiments, we first equilibrated the system at the initial temperature to obtain relative concentrations of all the species. With those initial concentrations, the temperature was then set to its final value and the kinetic master equation was solved to obtain the time-dependent population relaxation of each *n*-mer state. Similar types of models have been reported earlier for fitting experimental relaxation kinetics data.[23,25,27,28]

2.6 Coarse-grained AWSEM simulations The tethered WW-domains were computationally simulated using the Associative memory, Water mediated, Structure and Energy Model (AWSEM)[29]. The model predicts structures and helps understand the competition between folding

and inter-domain interactions by providing polymeric insights into the formation of contacts according to physico-chemical features of protein residues (sample structures in Figure 1).

AWSEM is a coarse-grained protein model with transferable energy functions that have been optimized to predict tertiary structures of globular proteins. AWSEM has been used in globular protein structure prediction, binding predictions of protein dimers, and amyloid fibril formation, through simulated annealing.[13,30–32] The AWSEM potential is a combination of both knowledge-based and physics-based terms. It uses a three-bead per amino-acid coarse-graining (C_α, C_β, and O atoms) that generates the coordinates of other heavy atoms along a backbone. It includes independent and cooperative hydrogen bonding, water-mediated tertiary interactions, and biasing local structural preferences based on short fragment memories. Although AWSEM lacks atomistic resolutions, this model is sufficient in sampling a wide conformational space involving folding and binding contributing to the folded, the unfolded, or the misfolded states. Our model uses the native state of an individual WW domain as a reference state, and thus conservatively assesses the population of misfolded states.

The total Hamiltonian consists of a backbone term $\mathcal{H}_{BB}$, a potential of mean force $\mathcal{H}_{PMF}$, and a fragment memory term $\mathcal{H}_{FM}$:

$$\mathcal{H}_{AWSEM} = \mathcal{H}_{BB} + \mathcal{H}_{PMF} + \mathcal{H}_{FM}. \quad (4)$$

$\mathcal{H}_{BB}$ constrains the backbone chain to physically realistic heteropolymer conformations (see Supplementary Information section 3 for details). The potential of mean force $\mathcal{H}_{PMF}$ depends on the identities of the interacting residues and contains direct contacts, water-mediated contacts, burial, and hydrogen bonding terms (see Supplementary Information section 3 for details). The $\mathcal{H}_{BB}$ and $\mathcal{H}_{PMF}$ terms do not depend on the knowledge of the native structure and only depend on the sequence of residues; thus, the two terms allow for non-native and long-sequence distance interactions and are responsible for the formation of non-native structure across multiple domains. Model parameters are chosen to minimize misfolding of the WW domain by itself, in accord with experimental observation (see section 3.6).

The fragment memory term $\mathcal{H}_{FM}$ is particularly important in the context of single domain folding, as it contains local sequence interactions using the knowledge of the native structure. Memories are sequences with known structures (typically obtained from the protein data bank). The fragment memory potential sums over all memories *m* from short sequences, and all pairs of atoms (not residues) *i* and *j* such that the atoms have a sequence separation $3 \leq |I - J| \leq 9$, having the form

$$\mathcal{H}_{FM} = -\lambda_{FM}\sum_m\sum_{i,j\ni 3\le|I-J|\le 9}\exp\left[-\frac{(r_{ij}-r_{ij}^m)^2}{2\sigma_{ij}^2}\right] \quad (5)$$

where r_{ij} and r_{ij}^m are the distances between atoms i and j of the simulated structure and of the memory structure, respectively. In this study, we use a single memory, which is the folded experimental structure of the isolated FiP35 WW domain (PDB ID: 2F21)[33]. The well width σ_{IJ} $=\lambda_\sigma|I\text{-}J|^{0.15}$, and we fixed λ_σ = 0.2 Å in all simulations. Unlike Gō or structure-based models[34], $\mathcal{H}_{FM}$ only acts less than 10 residues apart within the monomer, affecting mainly secondary structure. The other non-backbone terms act both locally and in long-sequence distances, affecting tertiary and interdomain structure also. A more detailed description of the AWSEM Hamiltonian terms can be found in refs. [29,35,36] or at http://awsem-md.org/index.html.

The fragment memory terms contain a scaling parameter λ_{FM} in eq. (5) adjusting the interaction strength. λ_{FM} allows us to tune the aggregation propensity relative to the folding propensity. In this study, we use three different values of λ_{FM} to compare folding *vs*. aggregation of the tethered domains as bias towards the folded crystal structure is decreased (λ_{FM} = 0.4 kJ/mole, Model I), or increased (λ_{FM} =1.2 kJ/mole, Model III) compared to the standard value (λ_{FM} = 0.8 kJ/mole, Model II). All other parameters were kept at default settings.[29] The full set of AWSEM parameters used is shown in Supplementary Information (Supplementary Information section 4).

To further correct the secondary structure bias, we used the Protein Secondary Structure Prediction server JPRED[37], which provides information to adjust terms in $\mathcal{H}_{PMF}$.[38] Details are in Supplementary Information section 3, and secondary structure prediction is shown in Supplementary Information section 4.

2.7 System preparation and simulated annealing protocol We built the single memory configuration using atomic coordinates provided in the WW-domain of the human FiP mutant crystal structure with PDB ID: 2F21. We matched the sequences of the WW-domain oligomers used in the experiments (Table S1 in Supplementary Information). Each individual domain in the oligomers used the single memory of the monomer, and linkers joining domains were not influenced by the fragment memory term.

We performed all simulations in the canonical ensemble (NVT) using the Nosé-Hoover thermostat implemented using the LAMMPS molecular dynamics software.[39] To predict the structures, we performed annealing simulations starting from a linear extended peptide structure at a temperature of 650 K, and slow cooled over 10 million time-steps to 300 K (where a time-step is approximately 5 fs). Initial velocities were chosen randomly from a Boltzmann distribution with

the average squared velocity equal to $3k_BT/m$, where k_B is the Boltzmann constant, m is the mass, and the temperature T is set equal to 650 K. The simulated annealing was repeated 40 times for each oligomer and the three λ_{FM} values (Models I, II, and III). The temperature range was chosen to be approximately 150 K above and below the folding temperature of the monomer.

2.8 Order Parameters The fraction $Q^{(d)}$ of native contacts of domain d ranges from 0 to 1, with a higher value corresponding to greater similarity to the native structure, which in this case is the monomer WW-domain crystal structure. It is defined as

$$Q^{(d)} = \frac{2}{(N-2)(N-3)} \sum_{i<(j-2)} \exp\left[-\frac{\left(r_{ij}^{d} - r_{ij}^{m}\right)^2}{2\sigma_{ij}^2} \right] \quad (6)$$

where N is the number of residues in a single domain, $\sigma_{ij} = |i-j|^{0.15}$Å, r_{ij}^{d} and r_{ij}^{m} are the distances between C_α atoms i and j of the simulated structure in domain d and of the single memory structure, respectively.

The misfolding parameter $Z_{mf}^{(d)}$ measures the degree of misfolding of domain d (where larger values positive mean more misfolding, and 0 means folded). It has the form:

$$Z_{mf}^{(d)} = \left[\sum_{k \neq d}^{\sum_{i,j \in \mathbb{I}_k^d} \Theta(\xi_1 - r_{ij})} \right] \cdot \Theta\left(\xi_2 - Q^{(d)}\right) \quad (7)$$

where $\Theta(x)$ is the Heaviside step function, ξ_1 = 6.0 Å and ξ_2 = 0.75 are distance and folding thresholds, k is the index of the domain interface $\mathbb{I}_k^d$ from domain d. The set $\mathbb{I}_k^d$ is defined as the set of residues i in domain d and the set of residues j in domain k, and r_{ij} is the distance between C_α atoms i and j. The left term in brackets in the order parameter $Z_{mf}^{(d)}$ quantifies the amount of interfacial contact (meaning shared contacts between domains), while the Heaviside function at the end of the equation prevents $Z_{mf}^{(d)}$ from counting complete folded structures as misfolded, even if a domain shares a large interfacial contact. Then to calculate the probability $P_m(m \geq \mu \,|\, n)$ that at least μ domains out of n domains in an oligomer are misfolded, we first find the probability that domain i misfolds by counting the number of annealed structures that have a value of $Z_{mf}^{(d)}$ greater than $N/2$, where N is the number of residues in a single domain:

$$p_i = \left\langle \Theta\left(Z_{mf}^{(i)} - N/2\right) \right\rangle \quad (8)$$

Therefore,

$$P_m(m \geq \mu | n) = \sum_{i_1 < \ldots < i_\mu}^{n} \prod_{k=1}^{\mu} p_{i_k} \quad (9)$$

where $n \geq \mu$, otherwise P_m=0. The total number of C_β to C_β contacts per domain is defined as:

$$\tau_{C\beta} = \frac{1}{n} \sum_{i<(j-2)} \Theta(r_{ij} - \xi_1). \quad (10)$$

Here $\tau_{C\beta}$ is the sum of all possible pairs of C_α atoms *i* and *j* that are closer than the cut-off distance ξ_1, divided by the number of domains in the oligomer, *n*.

As a final order parameter, the total number of Trp contacts per domain is defined as:

$$\tau_W = \frac{1}{n}\sum_{i \in Trp}\sum_{\substack{j \\ |i-j|>2}}\Theta(r_{ij} - \xi_1) \quad (11)$$

τ_W is the sum of all possible pairs of tryptophan "*i*" and other residues "*j*" closer than the cut-off distance ξ_1.

3. Results

3.1 Decrease in expression yield from monomer to pentamer One sign that an oligomer is prone to aggregation or self-aggregation is a decrease in expression yield of the protein. Other possibilities include protein toxicity (WW domain is well-tolerated in *E. coli*), protein length (the constructs are well below the 100 kDa length where we usually encounter expression problems for large fluorescent-labeled proteins), or increased degradation. We observed repeatedly (3-6 repeats) that as oligomer size was increased, the expression yield decreased.

The expression yield of the monomer was in the range of 8-12 mg per three-liter expression volume. The dimer also had a 8-12 mg yield range. The trimer had yield in the 5-7 mg range, slightly lower than monomer or dimer. The tetramer construct with a His tag yielded only 3-5 mg of protein, significantly lower. The tetramer solution also turned turbid as fractions were collected on the FPLC, which was not observed for the other proteins. For the tetramer, the purification tag had an important influence on expression yield. The tetramer using a GST tag[23] had minimal expression, although both versions of QFiP35, when characterized by MALDI (see Supplementary Information section 5) had a clear peak at 17.76 kDa for the cleaved QFiP35. The pentamer could not be expressed with measurable yield.

3.2 Changes in CD signal from monomer to tetramer The circular dichroism (CD) spectrum of proteins with His- or GST-tag removed confirms the expression trend. MFiP35 had a tryptophan chiral peak at 225 nm with mean residue ellipticity ~2000 deg $dmol^{-1}$ cm^{-2}, similar to previous literature reports (e.g. Figure 6 in ref. 40). Interestingly, the dimer and trimer show an even more intense peak, despite normalization to mean residue ellipticity in Figure 2. The environment of the Trp backbone must be even more chiral in these constructs than in the monomer, indicating some domain interactions even for well-folded and relatively high expressing oligomeric repeats (see also fluorescence results below). QFiP35 with a His tag again shows a smaller ellipticity, and QFiP

expressed with GST and then cleaved shows no significant peak in the CD spectrum between 210 and 260 nm. Each expression was repeated at least three times and similar results were obtained. The monomer showed no signal difference, whether expressed with His-tag or GST.

Clearly the tetramer is very sensitive to its local folding environment in *E. coli* during expression (His tag vs. GST tag), and folding is not successful with the GST tag. We surmise that the nascent QFiP35 chain during translation interacts with hydrophobic patches on the GST surface and misfolds, or that weak association of GST tags brings FiP35 repeats into proximity, whereas the His tag has no such effect in the bacterial cells used for protein expression.

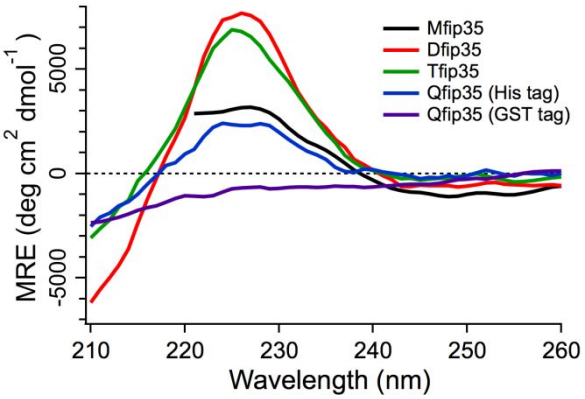

Fig. 2. Comparison of QFiP35 (tetramer) expressed and purified using GST and His tag using circular dichroism at 25 °C. The typical 227 nm peak for the WW domain is present in QFiP35 (His tag) protein but not in the spectra obtained for QFiP35 (GST tag).

3.3 Decrease of thermal stability from monomer to tetramer The thermal stability of the tethered *n*-mer constructs was measured by probing the only tryptophan (present in the hairpin 1 in each monomer WW Domain) over a temperature range of 5-90 °C by tryptophan fluorescence spectroscopy. The thermal melts were performed with varying concentrations of guanidine hydrochloride (0, 1, 2, 3 M) to obtain the melting temperature (T_m) with better accuracy by having more folded and unfolded baselines sampled for the global fit of all fluorescence data of all oligomers.

We make the following two observations in Figure 3. Fig, 3A-D shows a small decrease of the average Trp fluorescence wavelength (347 to 343 nm) in the native state as *n* goes from 1 to 4. This indicates a less polar environment of the tryptophan, consistent with some domain interaction reducing the tryptophan solvent exposure. Figure 3E directly compares the stability of all four constructs in 2 M GuHCl. The comparison is made at 2 M GuHCl because all but MFiP35 have reasonably complete native and denatured baselines at that denaturant concentration. MFiP35 denaturation does not reach the unfolded baseline ($T_m \geq 68$ °C for a fit of eq. (2) to that trace); it is the most stable protein. DFiP35 and TFiP35 do reach the unfolded baseline ($T_m = 63 \pm 1$ °C), and

show a slightly earlier onset of denaturation than the monomer. QFiP35 (His tag) is slightly less stable (T_m = 61±1 °C) than dimer and trimer. The decrease in stability is also supported by the global fitting model described further below. Thermal melts detected by CD yield melting temperatures about 4 °C lower than tryptophan fluorescence detection for MFiP35 though QFiP35 (His tag).

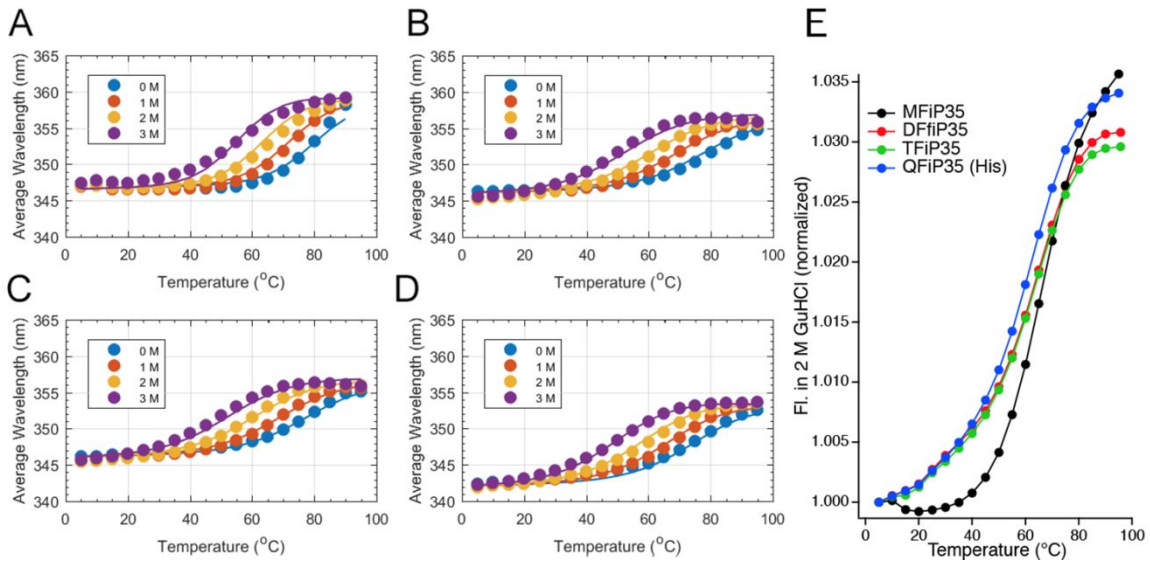

Fig. 3. Global fitting (A-D) of the thermal melts at different GuHCl concentrations and comparison at 2 M GuHCl (E). (A) MFiP35; (B) DFiP35; (C) TFiP(35); (D) QFiP35 expressed with a His tag. The curves are from the global multi-domain model fit of all thermodynamic and kinetic data of all *n*-mers, assuming three states per domain. (E) Comparison of the data at 2 M GuHCl, with connected data points to guide the eye T_m changes M>D,T>Q.

The difference between QFiP35 expressed with His and GST tags is seen in thermal melts, monitored in Figure 4 by tryptophan fluorescence wavelength shift (eq. (1)). Sigmoidal melts were fitted to the linear free energy model in eq. (2). In both proteins, Trp fluorescence starts at ~342 nm at low temperature, indicating partly buried Trp residues. Thermal denaturation of QFiP35 purified with the His tag shows a regular decrease of stability as GuHCl is added and unfolding to a solvent exposed state with an average wavelength of ~352 nm. In contrast, the GST-purified construct shows more erratic curves and sometimes even a wavelength decrease at high temperature. We believe the wavelength decrease indicates onset of inter-domain aggregation (tryptophans are re-introduced to a less polar environment at high temperature).

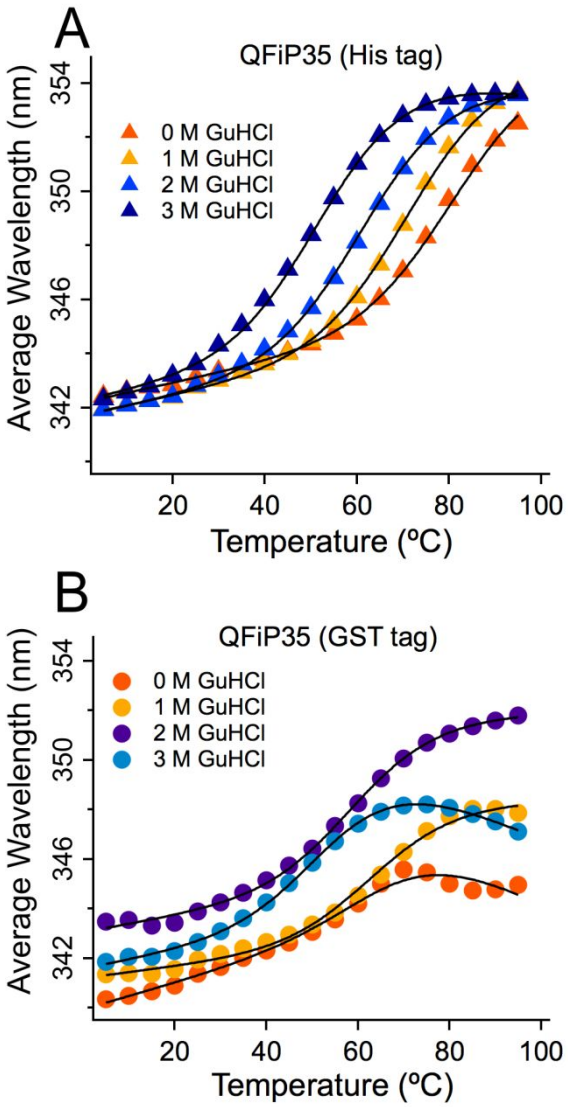

Fig. 4. Comparison of average Trp fluorescence wavelength *vs.* temperature plot for (A) QFiP35 (His tag, curves are global fit from Figure 2) and (B) QFiP35 (GST tag, curves are individual sigmoid fits to eq. 2). Melting temperature is higher and more regular sigmoid denaturation curves are obtained in the case of His-tag purification. Addition of GuHCl shifts the folded baseline in the GST-tagged protein.

3.4 T-jumps kinetics as a function of *T*, [GuHCl] and *n* In order to determine the unfolding relaxation kinetics, we conducted temperature jump experiments on all the tethered constructs except QFiP35 with GST tag. The jumps were conducted near and below the melting temperature using our ultrafast laser temperature jump setup described in the Methods section. The kinetics experiments were also measured at different temperatures and GuHCl concentrations for each *n*-mer (Figure 5 and Figures S6a-c in Supplementary Information section 6).

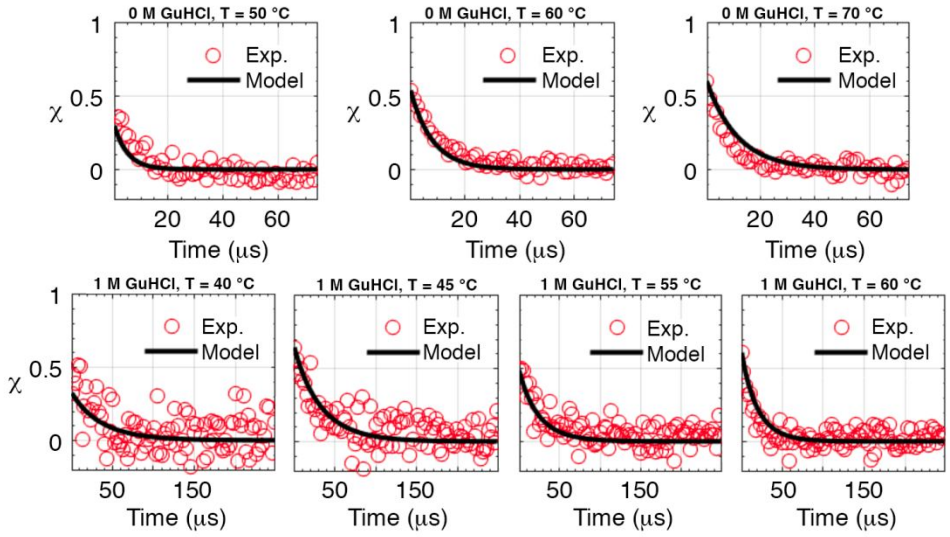

Fig. 5. Representative plots for temperature jump relaxation kinetics. χ(t) *vs.* time traces are fitted using the global fitting model for the monomer MFiP35 (top) and tetramer QFiP35 with histidine tag purification (bottom). The other proteins are shown in Supplementary Information section 6 in Figures S6a-b. The black curves are from a global fit of all thermodynamic and kinetic data of all *n*-mers simultaneously.

3.5 Global fitting model The thermal/GuHCl denaturation and T-jump kinetics data for all proteins with *n*=1-4 was fitted simultaneously using the model described in the section 2.5. Briefly, the model assumes that each protein can be represented as a chain "X" through "XXXX" for the monomer through tetramer,[9] where "X" stands for one of three states: N (native), U (unfolded) and M. The state M accounts for non-native structure of individual domains due to inter-domain interactions in the dimer through tetramer, not to be confused with short-lived two-stranded on-pathway intermediates that have been observed during WW monomer folding. Only direct interconversion between N and U, and U and M one step at a time was allowed (i.e. M is treated as "off pathway"). We also fitted an analogous model with only two states per domain, N and U, to assess the extent of domain interaction modeled by state M. The monomer can only access states N and U in both models.

Table 1: Global fitted parameters. The global fit incorporates all fluorescence-detected thermal melts and kinetics of all oligomers simultaneously. One standard deviation errors are shown in parentheses. Supplementary Information section 2 discusses each parameter in detail: One effective melting temperature, five thermal and denaturant linear free energy parameters, four fluorescence baseline parameters, and two activation free energy parameters. The parameters for state M are missing in the fit with only two states per domain, which has a much worse error (χ^2).

Parameters	**3 states per domain model**	**2 states per domain model**
T_m (°C)	82.4 (0.5)	82.0 (0.6)
$g^{(0)}_{UM}$ (J/mol^{-1})	213 (100)	-
g_{UN} (J mol^{-1} K^{-1})	377 (15)	290 (12)
g_{UM} (J mol^{-1} K^{-1})	163 (20)	-
m_{UN} (J/mol^{-1} molar^{-1})	3145(200)	2060 (138)
m_{UM} (J/mol^{-1} molar^{-1})	1140 (300)	-
b_U	358.7 (0.2)	358.1 (0.3)
a_U	-0.03 (0.02)	-0.10 (0.02)
b_F	347.1 (0.2)	346.4 (0.3)
a_F	0.011 (0.001)	0.008 (0.005)
$G^{\ddagger}_{NU}$ (J/mol^{-1})*	18500 (8000)	19900 (1300)
$G^{\ddagger}_{MU}$ (J/mol^{-1})*	2600 (fixed)	-
Avg. χ^2 (thermo)**	1.7	23
Avg. χ^2 (kinetics)**	1.5	3.1

* $G^{\ddagger}_{NU}$ and $G^{\ddagger}_{MU}$ are barrier height going from folded to unfolded or intermediate to unfolded form
** 16 thermal melts with 288 data points total and 30 kinetic decays each with 674 data points

The resulting network of states is described by 12 (8) thermodynamic parameters for three (two) states per domain (Table 1), as defined in Methods. The kinetic parameters include two activation free energies $G^{\ddagger}$ for the N to U and M to U reactions. (See also Supplementary Information section 2 for more details.)

Both the three-state and two-state models could globally fit all 46 denaturation and kinetic traces of all tethered oligomers n=1-4 simultaneously, as shown in Figures 3 and 5, and SI Figures S6a-h. The fitted model parameter values are shown in Table 1 for both global fits. It is evident by comparing Figures 3, 5 and S6a-b (three states per domain) with Figures S6c-g (two states per domain) that the state M, which accounts for domain interactions, is required for a satisfactory global fit of all data. The χ^2 of the global fit with three states per domain is 13 times smaller for equilibrium melts, and less than half for kinetics data, relative to the global fit with just two states

per domain. This trend is also observed if state "M" is excluded for the monomer, which has no experimental off-pathway intermediates.

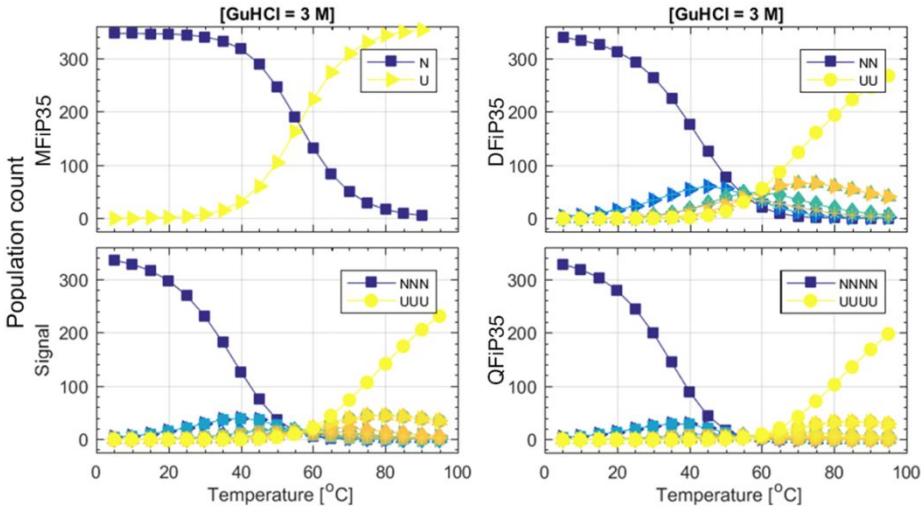

Fig. 6. Global fitting model population examples: As temperature increases at 3 M GuHCl, population of the fully native monomer (dark blue, N) decays and the unfolded monomer U builds up; there is no misfolded aggregate state for the monomer. For dimer through tetramer, non-native aggregate states containing 1 or more "M" (light blue, green, orange) build up, as does the fully unfolded state UU or UUUU (yellow). For a full plot of all *n*-mers and all GuHCl concentrations see Supplementary Information section 7.

Thermodynamic melts fitted to an effective T_m of ~84 °C for the monomer, with a single N and U baseline per monomer across the entire thermal melt data set (see Figure 3). The kinetic data was globally fitted with a ~17 kJ/mol^{-1} NU barrier and a smaller fixed UM barrier (as shown in Figure 5 and Supplementary Information section 7).

From the global fit we can extract the population of each state. Figure 6 illustrates the buildup and decay of all states as a function of temperature and GuHCl concentration. Fig. S7 shows the full data set. As expected, the fully native oligomer concentration decreases upon increased temperature or addition of denaturant. For longer oligomers, the fully native population decays more quickly in favor of oligomers containing some domains in the "M" state. The concentrations of all species can be recreated using Table 1 and the free energy formula eq. (3), showing that longer oligomeric tethered constructs have a higher propensity for occupying state M at high temperature. Our assumption that the fluorescence signature of monomers in state M is half-way between N and U, if relaxed, of course leads to different populations of misfolded states, but the general picture remains the same as shown in Figure 6. Although the model fit is quantitative, the populations in Fig. 6 should be taken only as a qualitative indicator.

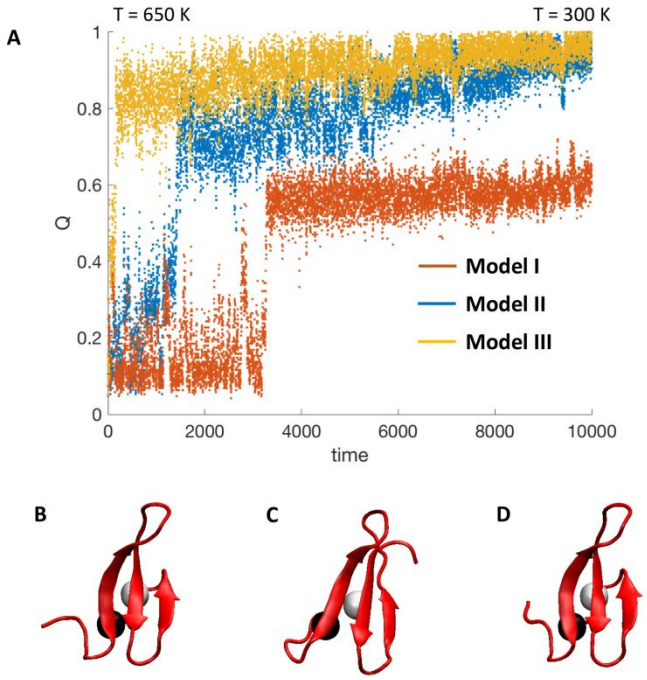

Fig. 7. (a) Simulated annealing trajectories with respect to fraction of native contacts Q for WW-domain monomer for model I, II, and III. Trajectories start at 650 K and are gradually cooled to 300 K. Time is represented in units of 10^3 timesteps. Below the trajectory are the annealed monomer structures, from left to right: Q = 0.72 for model I (b), Q = 0.95 for model II (c), and Q = 1.0 for model III (d). The black and white beads are the Cβ atoms of Trp 8 and Tyr 20, respectively.

3.6 Behavior of the monomer in the coarse-grained model To provide more detailed structural information, we used AWSEM to perform simulated annealing starting with unfolded structures at 650 K and gradually reduced the temperature to 300 K to sample increasingly folded structures. We first looked at the WW monomer. Figure 7 compares the simulated annealing trajectories for the three Models with λ_{FM} =0.4 kJ/mole (Model I), 0.8 kJ/mole (Model II) and 1.2 kJ/mole (Model III). As discussed in Methods, λ_{FM} defines the strength of the fragment memory Hamiltonian, with large values favoring folding over domain interactions. As λ_{FM} increases, the WW-domain collapses and folds earlier and at higher temperature. At $Q \approx 0.35$, only a single β-hairpin is formed (see an example in Figure 8B.). At $Q \approx 0.65$, all three β-strands form, but sidechains are not quite natively packed yet. At $Q \approx 0.95$, the protein is folded. Model I does not fold sufficiently well to be consistent with experiment, whereas Model II and III completely fold, consistent with fully native structure of the monomer. We favor Model II because it has less weighting on the fragment memory interactions (smaller value of λ_{FM}) than Model III, thus achieving complete folding of the monomer without over-weighting native interactions.

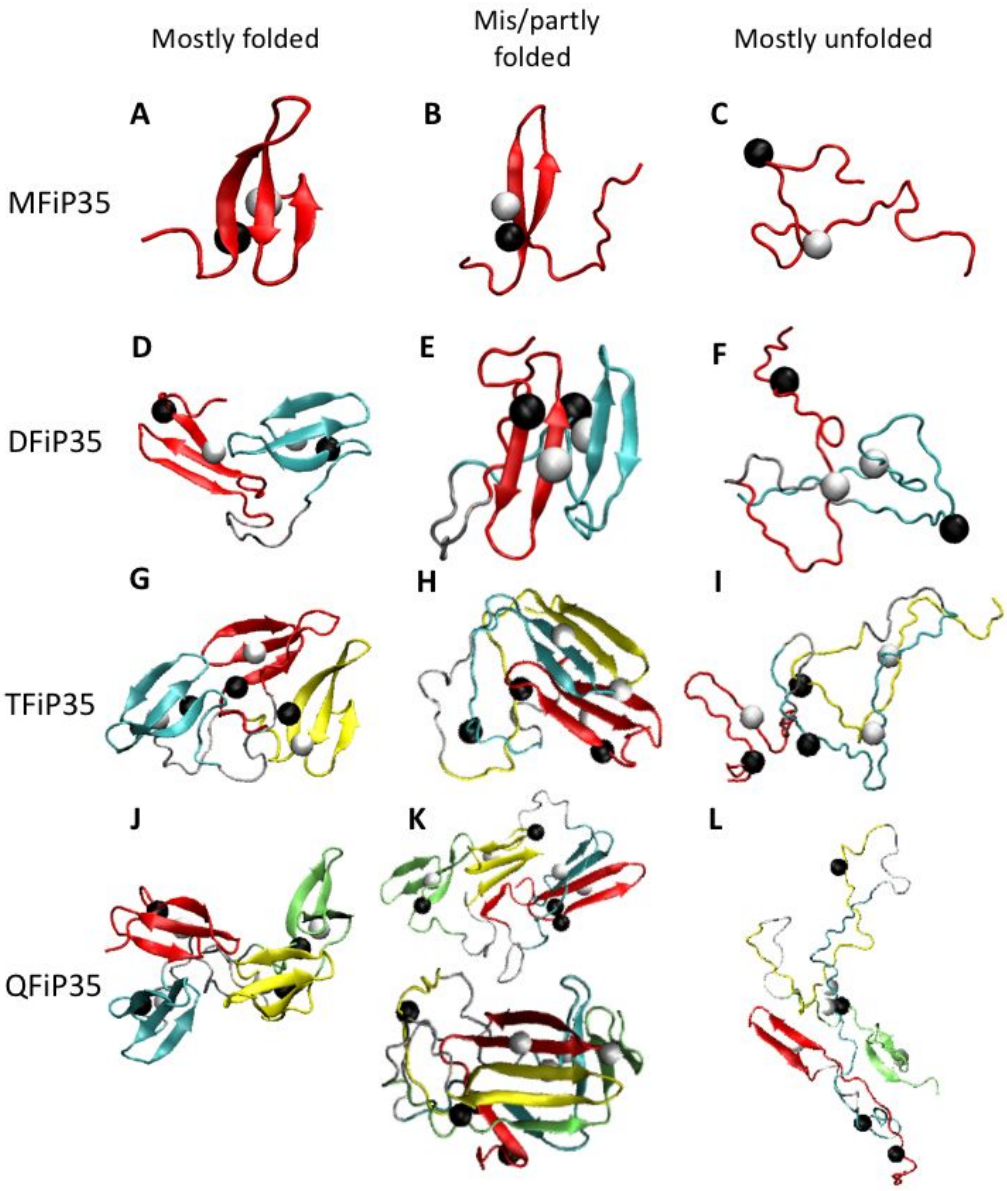

Fig. 8. Gallery of oligomers. Examples of predicted WW-domain monomer, dimer, trimer, and tetramer structures (from top to bottom) with varying amounts of folding/misfolding of domains. The domains are colored red-cyan-yellow-green from the N-terminus, and linkers are gray. The black and white beads are the C_β atoms of Trp 8 (or 51, 94, 137) and Tyr 20 (or 63, 109, 152), respectively. Oligomers states corresponding roughly to the discrete global fitting model for the experimental data: (A) N, (B) M, (C) U, (D) NN, (E) MM, (F) UU, (G) NNN, (H) NMM, (I) UUU, (J) NNNN, (K) NNMM and MMMM, (L) UUUU.

3.7 Coarse-grained simulations and structural interpretation of the data Simulations were next performed on tethered repeat proteins to get a higher resolution picture of the type of misfolded structures that may form. In agreement with experiment, simulation revealed that as the number of domains increases, the probability of misfolding increases. A gallery of monomers and oligomers with varying degrees of folding/misfolding from simulated annealing is shown in Figure

8, and annealing results are shown in Supplementary Information section 8. As the number of domains increases (MFiP35 to QFiP35), the folding of the individual domains competed less effectively with inter-domain interactions. A common feature of the misfolded structures is that one β-strand unfolds, and the remaining two β-strands from separate domains come together to make a larger β-sheet (e.g. structures 8E and 8H). This misfolding mechanism is clearly seen in the annealing trajectory of a trimer (Fig. S8b). The tetramer exhibited an additional type of misfolding (structure 8K) by forming chimeric β-sheets with domain-swapped structures. Thermodynamically, the non-fragment memory terms of the Hamiltonian in eq. (4), primarily the Ramachandran term, $\mathcal{H}_{rama}$ (see SI eq. (S5)), the β-strand hydrogen bonding term, $\mathcal{H}_{\beta}$ (see SI eq. (S6)), and parallel-antiparallel cooperative hydrogen bonding term, $\mathcal{H}_{P-AP}$ (see SI eq. (S7)), are responsible the formation of the β-sheets across multiple domains.

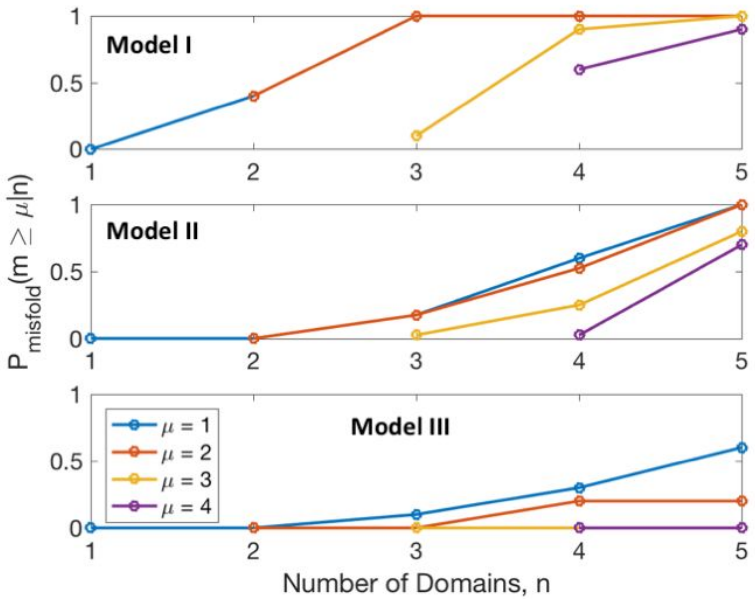

Fig. 9. Probability $P_{misfold}(m \geq \mu|n)$ of misfolding μ or more domains, given the size of the n-mer from n = 1 to 4. Models I, II, and III are shown from top to bottom. Probabilities are calculated from structure predictions of simulated annealing runs.

3.8 Misfolding propensity increases with oligomer size The experimental expression yield trends are supported by coarse-grained simulations of the tethered systems. Figure 9 shows the probabilities for μ or more domains in the n-mer to misfold, or $P_{misfold}(m \geq \mu|n)$, for models I, II, and III. The scaling factor λ_{FM} controls the bias towards the monomer native structure, with smaller values leading to more interaction among domains. For smaller λ_{FM} (model I), $P_{misfold}(m \geq \mu|n)$ is driven towards 1 for smaller repeat proteins. The probability of misfolding of at least one domain is >0.5 for the tetramer in model II, which fits well with the observed intracellular

environmental sensitivity of the tetramer as seen by decrease in yield and sensitivity to the type of purification tag being attached. Model II shows no significant effect on monomer and dimer, consistent with the onset of lower melting temperature for the trimer in the thermal melts performed on the tethered proteins (Table 1). Even in model III, which has the strongest domain folding propensity, the tetramer has at least one domain misfolded with a probability of 0.3.

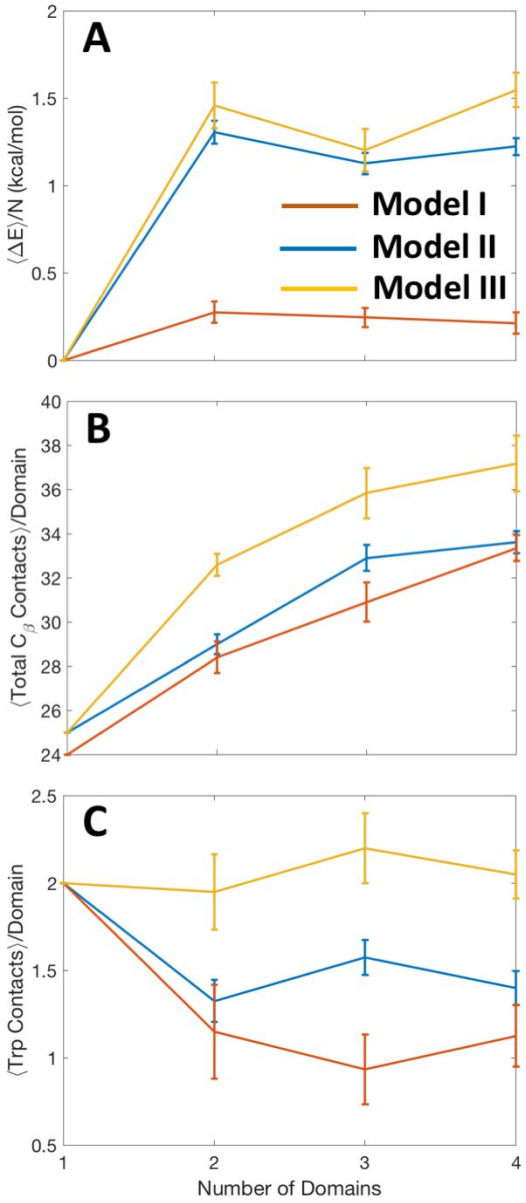

Fig. 10. (a) Energy change (from monomer) per residue vs. number of tethered domains. (b) Total number of C_β to C_β contacts per domain vs. number of domains. (c) Number of Trp contacts per domain vs. number of domains, for model I in orange, model II in blue, and model III in yellow.

With increasing number of domains, the probability of misfolding increases due to increased competition of inter-domain interactions with folding, shifting equilibrium towards misfolded states. Another possible reason is that because not all the β-strands form at the same time,

misfolding can occur when β-sheets of neighboring domains interact and become kinetically trapped beyond the time scale of the experiments. The gallery of *n*-mers in Figure 8 is consistent with both scenarios, although we favor the equilibrium scenario for two reasons: in the model, extensive simulated annealing was applied; and in the global fitting model (Figure 6), equilibrium is achieved while accurately fitting the experimental data. Furthermore, the strong coupling between domains reflected by $P_{misfold}$ increasing with *n* (Figure 9) and the mis/partly folded structures in Figure 8 validate the global fitting model assumptions (section 2.5) and fitting results (section 3.5). It is reasonable for state M to represent non-native structure of a domain due to domain interactions, rather than an isolated misfolded state of WW domain.

3.9 Order parameters track Trp fluorescence, stability and misfolding Figure 10 presents the averages of three order parameters with respect to number of domains for models I, II, III. The number of contacts made by tryptophan (Figure 10C) can be correlated with the fluorescence spectroscopy, since fewer contacts imply more solvent exposure and red-shifted fluorescence. The number of Trp contacts is highest for the monomer signifying a stable native structure. This is also verified with the change in energy per domain (Figure 10A), which shows the monomer as the most stable compared to the other oligomers. Figure 10B shows that more C_β contacts form as the number of domains increases, signifying an increase in hydrogen bonding between β stands of different domains. This increase in C_β contacts can be visualized as an increase in chimeric β-sheets seen in Figure 8E, H and K. This analysis is consistent with the experimental results obtained by CD and fluorescence spectroscopy.

3.10 Local backbone geometry remains conserved while global configuration depends on *n* The CD spectra in Figure 2 vary in intensity, but generally have the same shape, indicating similar local backbone configurations. Figure 11 plots Φ and Ψ Ramachandran angle probabilities for the different *n*-mers. Even though there is a clear change in the structures globally as more domains are added (Figure 8), the local secondary structure landscape is preserved in Figure 11. As expected from the high amount of β-sheet formation (either within a single domain or across multiple), the most probably angles are those that are prone to form β-sheets. The Ramachandran histogram also populates angles that have a high propensity of forming α-helices ($-70° > \Psi > +20°$) even though none of the proposed structures (*n* = 1 to 4), in Figure 7 or 8, contain an α-helix.

However, the angles with helical tendencies lead to an actual α-helix only in the coarse-grained simulated annealing of the pentamer (*n* = 5) in Figure 12, which could not be expressed in experiments. The pentamer forms a new type of misfolded structure compared to the ones seen in

Figure 8 for *n*=2-4: an α-helix containing a Trp residue in the second domain, which is surrounded by β-sheets, emerges in ~20% of predicted pentamer structures. This suggests that the extra domains stabilize an α-helix formed by residues with Ψ angles that are in the range $-70° > \Psi > +20°$. The extra domains provide more tertiary contacts allowing for side-chains to align correctly into an α-helix from an already twisted β-strand.

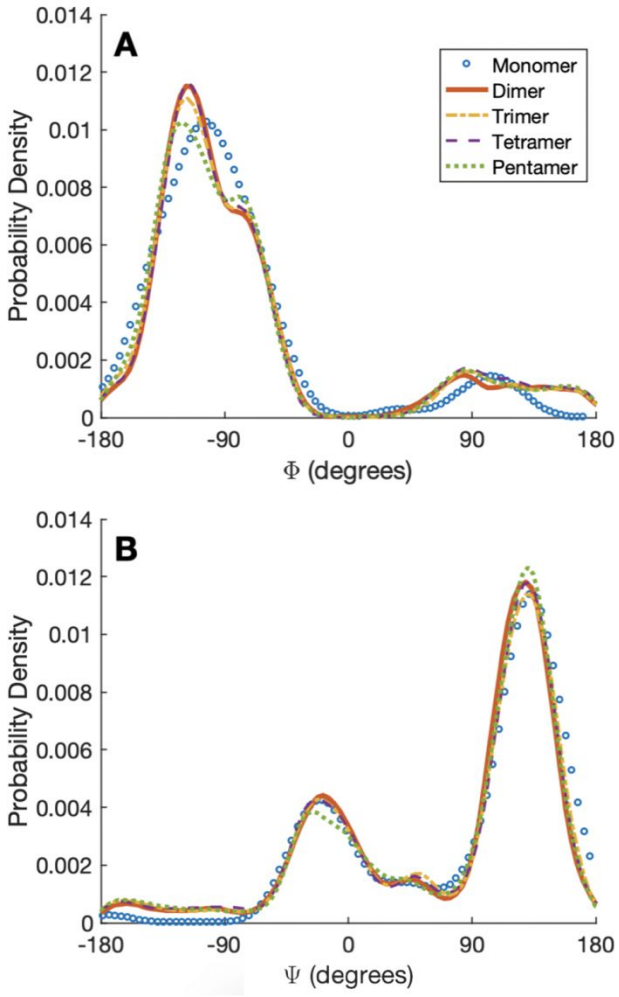

Fig. 11. Probability density of (a) Φ and (b) Ψ angles for monomer, dimer, trimer, tetramer, and pentamer for model II.

Additionally, the pentamer has a probability very close to 1 of at least one domain being misfolded, and 0.6 even for the conservative model III (Figure 9). The lack of pentamer expression again suggests that Model II represents a fairly accurate balance between fragment memory and inter-domain interactions.

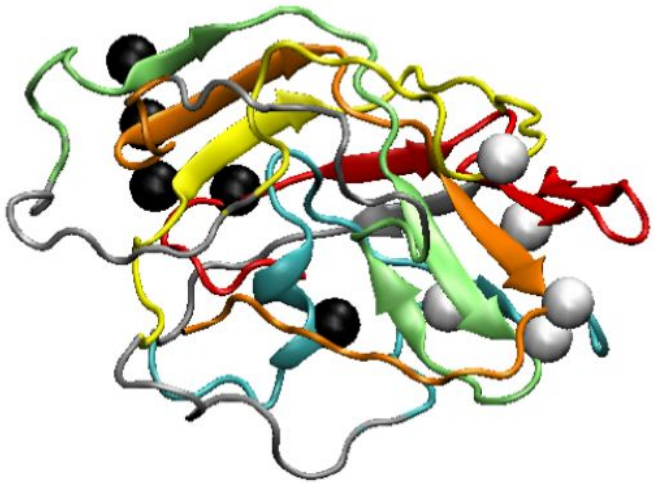

Fig. 12. Example of a predicted WW-domain pentamer structure with all domains misfolded. Domains are colored, and linkers are gray. The order of the colors starting from N-terminus is red, cyan, yellow, green, orange. The black and white beads are the C_β atoms of Trp (8, 51, 94, 137, 180) and Tyr (20, 63, 109, 152, 195), respectively. An α-helix containing a Trp residue appears in the cyan domain.

4. Discussion

Natural and engineered repeat proteins have provided many insights to relate folding, misfolding and function. Evolution for folding, which does not favor repeats with similar sequences adjacent in a multi-domain protein,[6,7,15] goes hand-in-hand with evolution for function, which sometimes favors multiple domains of similar structure.[21,41] Ankyrin domains[42,43] and TPR motifs[44] in particular have shown how nearly identical folds can co-exist with the right balance of sequence similarity. These results have been complemented by studies on consensus repeat sequences,[8,45] which have shown evidence of a highly parallel, but not completely homogeneous, folding process capable of generating stable native states.[3,9,46]

Our results for repeats of sequence-identical WW domains show that above n=3, a critical number of repeats is reached: individual domains are destabilized and likely to form non-native states. While the stability of dimer and trimer is at most slightly smaller than that of the monomer, the tetramer is noticeably less stable thermodynamically and sensitive to the purification tag used, whereas the pentamer cannot be expressed in significant quantities, presumably due to domain interactions that lead to misfolding. Thus, 2 to 3 identical repeat domains lead to a stable native WW repeat protein, but more identical domains in series foster misfolding. The observation that consensus ankyrin sequences (α-helical secondary structure) can form longer repeat folds than WW domain (β-sheet secondary structure) indicates that certain folds and sequences have substantially lower propensity for misfolding than others when tethered together. This may be the

reason why WW domains are mainly observed as tandem repeats in nature,[21] whereas natural ankyrins can contain many additional repeats.[47]

Transient aggregates have been proposed as a step during the folding of many non-repeat proteins, masquerading as folding intermediates. For example, the RNA-binding protein U1A forms such transient aggregates.[2] We have shown that when U1A is tethered into a repeat protein, transient aggregation is enhanced and leads to irreversible (on the time scale of the experiments) aggregation when too many repeats are tethered together.[47] U1A is a very aggregation-prone protein, and we found that the size of its irreversible aggregation nucleus is only n=2.[10] WW domain is not prone to aggregation (as evidenced by facile NMR structures obtained at mM concentration).[33] Here we find that the size of the FiP35 irreversible aggregation nucleus lies at n = 4. Thus, if a range of $n \approx 2$ to 4 is likely for the aggregation nuclei of most proteins; oligomeric aggregates may be formed rather easily. The 'intramolecular amyloids' we observe when repeats interact (e.g. Figure 8H) may be examples of what oligomeric aggregates in non-tethered proteins look like. Indeed, it has been shown for protein U1A that addition of an Alzheimer sequence increases transient aggregation and allows stable dimers to form.[48]

The WW tetramer highlights how protein folding can be sensitive to the environment, in tandem with current in-cell folding experiments.[49,50] The type of purification tag used for WW tetramer (histidine *vs.* GST) determines whether a native-like or a non-native secondary structure is recovered. Thus, the local environment is critical for the folding of the tetramer. In-cell experiments have shown that proteins can be stabilized or destabilized in the cellular environment, depending on protein surface electrostatics,[51] or the organelle environment.[52,53] While these effects are small, they can be critical in regulating signaling and other protein-protein interactions, which are often weak (on the order of a few kJ/mole).[1] Such sticking or 'quinary structure' of proteins,[54–56] of which only the tip of the iceberg has been characterized,[57,58] may well account for the large majority of in-cell protein-protein or protein-nucleic acid interactions.

Repeat proteins may assist in the evolution of new folds.[59–62] Our structural simulations of identical repeats highlight one possible path towards the evolution of more complex protein folds. For example the tetramer (Figures 8K) forms larger-stranded beta sheets by combining strands from different domains. Since the WW-domain monomer contains three beta-strands with strong curvature (seen in Figure 7), the beta-strands that form the disordered loops in the oligomers are prone to form helices. Such loops could evolve to form helical structure (Figure 11), yielding a protein whose beta sheets have large contact order[63] because they are separated by other secondary structure elements (loops, helices). The latter is a very common structural motif. Indeed, longer

repeats can form entirely novel structures, such as the one shown in Figure 12. Although the pentamer has many disordered regions, the combination of beta sheets and an alpha helix showed up in ~20% of simulated structures. If the loops were optimized by shortening, or mutated to favor additional alpha helices, Figure 12 would represent a compact alpha/beta fold. Although not the subject of this paper, it would be interesting to take a sequence that forms simulated compact misfolded structure such as in Figure 12, truncate the loops or increase their helix propensity, and see if improved expression and a well-defined tertiary structure could be obtained.

Acknowledgements

We thank Dr. Nicholas P. Schafer for the helpful discussions. K.D. and M.G. were supported by NIH grant R01 GM093318. A.G.G was supported by a training fellowship on the Houston Area Molecular Biophysics Program (T32 GM008280). M.S.C. and A.G.G. were funded by the National Science Foundation (MCB-1412532, PHY-1427654, OAC-1531814), and thank computing resources from the Center for Advanced Computing and Data Systems at UH.

References

1 A. J. Wirth and M. Gruebele, *BioEssays*, 2013, **35**, 984–993.

2 M. Silow and M. Oliveberg, *Proc. Natl. Acad. Sci.*, 1997, **94**, 6084–6086.

3 D. U. Ferreiro, C. F. Cervantes, S. M. E. Truhlar, S. S. Cho, P. G. Wolynes and E. A. Komives, *J. Mol. Biol.*, 2007, **365**, 1201–1216.

4 M. E. Zweifel and D. Barrick, *Biochemistry*, 2001, **40**, 14357–14367.

5 C. H. Croy, S. Bergqvist, T. Huxford, G. Ghosh and E. A. Komives, *Protein Sci.*, 2004, **13**, 1767–77.

6 M. B. Borgia, A. Borgia, R. B. Best, A. Steward, D. Nettels, B. Wunderlich, B. Schuler and J. Clarke, *Nature*, 2011, **474**, 662.

7 C. F. Wright, S. A. Teichmann, J. Clarke and C. M. Dobson, *Nature*, 2005, **438**, 878.

8 E. R. G. Main, S. E. Jackson and L. Regan, *Curr. Opin. Struct. Biol.*, 2003, **13**, 482–489.

9 T. Aksel and D. Barrick, *Biophys. J.*, 2014, **107**, 220–232.

10 F. Liu and M. Gruebele, *J. Phys. Chem. Lett.*, 2010, **1**, 16–19.

11 W. Y. Yang and M. Gruebele, *Biophys. J.*, 2006, **90**, 2930–2937.

12 Y. Javadi and E. R. G. Main, *Proc. Natl. Acad. Sci.*, 2009, **106**, 17383–17388.

13 W. Zheng, N. P. Schafer and P. G. Wolynes, *Proc. Natl. Acad. Sci.*, 2013, **110**, 1680–1685.

14 D. U. Ferreiro, E. A. Komives and P. G. Wolynes, *Q. Rev. Biophys.*, 2014, **47**, 285–363.

15 A. Borgia, K. R. Kemplen, M. B. Borgia, A. Soranno, S. Shammas, B. Wunderlich, D. Nettels, R. B. Best, J. Clarke and B. Schuler, *Nat. Commun.*, 2015, **6**, 8861.

16 M. B. Prigozhin and M. Gruebele, *J. Am. Chem. Soc.*, 2011, **133**, 19338–19341.

17 J. K. Weber, R. L. Jack, C. R. Schwantes and V. S. Pande, *Biophys. J.*, 2014, **107**, 974–82.

18 P. Tian and R. B. Best, *PLoS Comput. Biol.*, 2016, **12**, e1004933.

19 M. Faccin, P. Bruscolini and A. Pelizzola, *J. Chem. Phys.*, 2011, **134**, 075102.

20 T. Kajander, A. L. Cortajarena, E. R. G. Main, S. G. J. Mochrie and L. Regan, *J. Am. Chem. Soc.*, 2005, **127**, 10188–10190.

21 E. J. Dodson, V. Fishbain-Yoskovitz, S. Rotem-Bamberger and O. Schueler-Furman, *Exp. Biol. Med.*, 2015, **240**, 351–360.

22 X. Chen, J. L. Zaro and W.-C. Shen, *Adv. Drug Deliv. Rev.*, 2013, **65**, 1357–1369.

23 F. Liu, M. Nakaema and M. Gruebele, *J. Chem. Phys.*, 2009, **131**, 0–9.

24 E. R. Henry, R. B. Best and W. A. Eaton, *Proc. Natl. Acad. Sci.*, 2013, **110**, 17880–17885.

25 J. Ervin, J. Sabelko and M. Gruebele, *J. Photochem. Photobiol.*, 2000, **B54**, 1–15.

26 R. M. Ballew, J. Sabelko, C. Reiner and M. Gruebele, *Rev. Sci. Instrum.*, 1996, **67**, 3694–3699.

27 A. J. Wirth, Y. Liu, M. B. Prigozhin and K. Schulten, *J. Am. Chem. Soc.*, 2015, **137**, 7152--7159.

28 C. M. Dobson, in *Seminars in Cell and Developmental Biology*, 2004, vol. 15, pp. 3–16.

29 A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes and G. A. Papoian, *J. Phys. Chem. B*, 2012, **116**, 8494–8503.

30 W. Zheng, N. P. Schafer, A. Davtyan, G. A. Papoian and P. G. Wolynes, *Proc. Natl. Acad. Sci.*, 2012, **109**, 19244–19249.

31 W. Zheng, N. P. Schafer and P. G. Wolynes, *Proc. Natl. Acad. Sci.*, 2013, **110**, 20515–20520.

32 M. Chen and P. G. Wolynes, *Proc. Natl. Acad. Sci.*, 2017, **114**, 4406–4411.

33 M. Jäger, Y. Zhang, J. Bieschke, H. Nguyen, M. Dendle, M. E. Bowman, J. P. Noel, M. Gruebele and J. W. Kelly, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 10648–53.

34 C. Clementi, H. Nymeyer and J. N. Onuchic, *J. Mol. Biol.*, 2000, **298**, 937–953.

35 N. P. Schafer, B. L. Kim, W. Zheng and P. G. Wolynes, *Isr. J. Chem.*, 2014, **54**, 1311–1337.

36 G. A. Papoian, *Coarse-grained Modeling of Biomolecules*, CRC Press, 2017.

37 J. A. Cuff, M. E. Clamp, A. S. Siddiqui, M. Finlay and G. J. Barton, *Bioinformatics*, 1998, **14**, 892–893.

38 H. H. Truong, B. L. Kim, N. P. Schafer and P. G. Wolynes, *J. Chem. Phys.*, 2013, **139**, 121908.

39 S. Plimpton, *J. Comput. Phys.*, 1995, **117**, 1–19.

40 E. K. Koepf, H. M. Petrassi, M. Sudol and J. W. Kelly, *Protein Sci.*, 1999, **8**, 841–853.

41 P. Michaely and V. Bennett, *J. Biol. Chem.*, 1993, **268**, 22703–22709.

42 H. K. Binz, M. T. Stumpp, P. Forrer, P. Amstutz and A. Plückthun, *J. Mol. Biol.*, 2003, **332**, 489–503.

43 D. U. Ferreiro, S. S. Cho, E. A. Komives and P. G. Wolynes, *J. Mol. Biol.*, 2005, **354**, 679–692.

44 T. J. Magliery and L. Regan, *J. Mol. Biol.*, 2004, **343**, 731–45.

45 K. W. Tripp and D. Barrick, *J. Mol. Biol.*, 2007, **365**, 1187–200.

46 P. J. E. Rowling, E. M. Sivertsson, A. Perez-Riba, E. R. G. Main and L. S. Itzhaki, *Biochem. Soc. Trans.*, 2015, **43**, 881–888.

47 M. D. Jacobs and S. C. Harrison, *Cell*, 1998, **95**, 749–758.

48 D. E. Otzen, S. Miron, M. Akke and M. Oliveberg, *Biochemistry*, 2004, **43**, 12964–12978.

49 J. Danielsson, X. Mu, L. Lang, H. Wang, A. Binolfi, F.-X. Theillet, B. Bekei, D. T. Logan, P. Selenko, H. Wennerström and M. Oliveberg, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 12402–7.

50 I. Guzman, H. Gelman, J. Tai and M. Gruebele, *J. Mol. Biol.*, 2014, **426**, 11–20.

51 J. Danielsson, X. Mu, L. Lang, H. Wang, A. Binolfi, F.-X. Theillet, B. Bekei, D. T. Logan, P. Selenko, H. Wennerström and M. Oliveberg, *Proc. Natl. Acad. Sci.*, 2015, **112**, 12402–12407.

52 J. Tai, K. Dave, V. Hahn, I. Guzman and M. Gruebele, *FEBS Lett.*, 2016, **590**, 1409–1416.

53 A. Dhar, K. Girdhar, D. Singh, H. Gelman, S. Ebbinghaus and M. Gruebele, *Biophys. J.*, 2011, **101**, 421–30.

54 E. H. McConkey, *Proc. Natl. Acad. Sci. U. S. A.*, 1982, **79**, 3236–40.

55 K. S. Hingorani and L. M. Gierasch, *Curr. Opin. Struct. Biol.*, 2014, **24**, 81–90.

56 D. Gnutt and S. Ebbinghaus, *Biol. Chem.*, 2016, **397**, 37–44.

57 S. Sukenik, P. Ren and M. Gruebele, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 6776–6781.

58 C. Vélot, M. B. Mixon, M. Teige and P. A. Srere, *Biochemistry*, 1997, **36**, 14271–14276.

59 J. Söding and A. N. Lupas, *BioEssays*, 2003, **25**, 837–846.

60 C. A. Orengo, I. Sillitoe, G. Reeves and F. M. G. Pearl, *J. Struct. Biol.*, 2001, **134**, 145–165.

61 M. A. Andrade, C. Perez-Iratxeta and C. P. Ponting, *J. Struct. Biol.*, 2001, **134**, 117–131.

62 A. K. Björklund, D. Ekman, S. Light, J. Frey-Skött and A. Elofsson, *J. Mol. Biol.*, 2005, **353**, 911–23.

63 K. W. Plaxco, K. T. Simons and D. Baker, *J. Mol. Biol.*, 1998, **277**, 985–994.