Predicting unfolding thermodynamics and stable intermediates for alanine-rich helical peptides
with the aid of coarse-grained molecular simulation

Cesar Calero-Rubio^{1,⊥}, Bradford Paik^{2, ⊥}, Xinqiao Jia², Kristi L. Kiick^{2,}, Christopher J.*

Roberts^{1,}*

¹ *Chemical & Biomolecular Engineering Department, University of Delaware, Newark, DE 19716*

² *Material Science & Engineering Department, University of Delaware, Newark, DE 19716*

[⊥] Authors contributed equally

**Corresponding authors: (KLK) tel., 302-831-0201; email, kiick@udel.edu; (CJR) tel., 302-831-0838; email, cjr@udel.edu*

Keywords: coarse-grained modeling, protein unfolding, helical peptides, molecular dynamics, replica exchange, conformational stability

Abstract:

This report focuses on the molecular-level processes and thermodynamics of unfolding of a series of helical peptides using a coarse-grained (CG) molecular model. The CG model was refined to capture thermodynamics and structural changes as a function of temperature for a set of published peptide sequences. Circular dichroism spectroscopy (CD) was used to experimentally monitor the temperature-dependent conformational changes and stability of published peptides and new sequences introduced here. The model predictions were quantitatively or semi-quantitatively accurate in all cases. The simulations and CD results showed that, as expected, in most cases the unfolding of helical peptides is well described by a simply 2-state model, and conformational stability increased with increased length of the helices. A notable exception in a 19-residue helix was when two Ala residues were each replaced with Phe. This stabilized a partly unfolded intermediate state via hydrophobic contacts, and also promoted aggregates at higher peptide concentrations.

Introduction:

Unfolding, self-assembly and non-native protein aggregation play important roles in the solution behavior of proteins and polypeptides. In some cases, nonnative aggregation and self-assembly are undesired because they are responsible for the degradation of biopharmaceuticals and are linked to a number of debilitating diseases [1–3]. On the other hand, proper protein self-assembly is responsible for much of the physiological functions of proteins *in vivo*. [1,4–6]

Additionally, (poly)peptide self-assembly and aggregation can be used as a kinetically controlled and tunable process to form new structures based on various peptide sequences [7–9]. In each case, a molecular-scale description of the process(es) is needed if one wishes to design or predict the behavior and relative stability of key intermediate species – e.g., as a function of peptide sequence and solution environment [7,10–12]. This is experimentally challenging, as few experimental techniques allow for the identification of the role of each specific residue in the unfolding and aggregation behavior of a defined sequence [1,13]. Additionally, such techniques are relatively low throughput, expensive, and/or have large sample material requirements. This poses challenges for testing and design of a range of protein sequences and solution environments, and helps to motivate development of modeling approaches to aid in those efforts [10,14].

With the exception of relatively small proteins and peptides, it is too computationally expensive to simulate most systems of interest with all-atom simulations and explicit solvent molecules [15–18]. Atomistic force fields provide the most accurate method available to identify amino-acid-specific interactions, while coarse-grained (CG) molecular models provide faster computations at the expense of molecular definition [18,19]. Therefore, CG models can potentially be well-suited as a computational framework to speed up experimental searches in

order to optimize time and resources, provided the force-field can accurately predict the experimental behavior of interest [18,20,21].

In this regard, a number of 4 beads-per-amino acid (4bAA) CG models have been previously proposed and studied in order to capture the configurational and aggregation behavior of peptides and proteins in solution [7,16,22]. 4bAA models have primarily been designed to capture qualitative structural features of polypeptide unfolding and self-assembly – e.g., the transition between helix and coil configurations for natively helical polypeptides, the existence of local conformations during unfolding, and the formation of tertiary structures for long polypeptides [7,16,22]. Previous work from some of the present authors showed that 4bAA models can help one understand the molecular scale interactions that affect the unfolding and self-assembly of Ala-rich peptides in the context of polymer-peptide interactions [7]. An earlier model was based on that from Bereau and Deserno, which was originally designed to capture structural details of membrane protein assembly [16]. Separately, previous work has shown that the 4bAA level of CG modeling provides an optimal balance between molecular detail, computational accuracy, and computational burden when (poly)peptide unfolding and aggregation are of interest [22].

However, much less effort has been devoted to quantifying the thermodynamics of unfolding and aggregation in CG molecular models. Even though qualitative structural features of the unfolding process are reasonably well captured, the quantitative details such as unfolding free energy values, midpoint unfolding temperatures (T_m), and unfolding enthalpy values obtained from molecular simulations (both CG and atomistic) typically do not match those obtained experimentally [7,16,22–25]. Furthermore, it is common to use molecular simulations in a “hindsight” manner, where the experimental behavior is already known and the simulations

are intended to give molecular-scale insight that is beyond the capabilities of the experiment, or to help confirm or refute hypotheses that were based on interpretation of the experimental data. Much less work has been devoted to predicting experimental behavior *a priori* with CG models.

This motivates the first part of the present work, which is focused on refining the previously proposed 4bAA protein model [7,16] to more accurately quantify unfolding thermodynamics for a series of helical Ala-rich polypeptides as a function of chain length. From there, the model was used with Replica Exchange Molecular Dynamics to make *a priori* predictions for the unfolding thermodynamics and pathways for a set of new alanine-rich peptides that were then experimentally synthesized and characterized with circular dichroism spectroscopy, for comparison to model predictions. The particular choices of peptides were based on previous alanine-rich sequences and future applications that focus on control of peptide-peptide interactions in multiblock peptide-polymer conjugates to manipulate assembly [8,9,26–33]. The simulated and experimental unfolding thermodynamics were in quantitative or semi-quantitative agreement across the range of peptides tested, but the model also predicted outliers compared to the expected peptide behavior that was based on qualitative reasoning regarding amino acid hydrophobicity and the length of the peptide chain. In that case, stable folding intermediates were responsible for non-2-state behavior, and the CG simulations suggested key intra-molecular interactions that stabilized partly folded intermediate structures and complicated the unfolding transitions of even these relatively small, alanine-rich helices. Inter-peptide interactions were also shown to stabilize partly unfolding conformers that may be important as intermediates for larger-scale aggregation, consistent with preliminary experimental results.

Material and Methods

Peptide synthesis and purification

All materials were purchased from Fischer Scientific (Pittsburgh, PA) except where otherwise indicated. Peptides were synthesized on a Rink Amide Resin (ChemPep, Wellington, FL). Specifically, the sequences listed in Table 1 (shorthand notations used in Table 1: AQEK, FAQEK, and AQK18) were synthesized with a PS3 peptide synthesizer (Protein Technologies, Tucson, AZ). Longer sequences (shorthand notation in Table 1: AQK27 and AQK35) were synthesized with a Focus XC peptide synthesizer (AAPTec Inc, Louisville, KY). Fmoc-alanine, Fmoc-lysine(boc), Fmoc-glutamic acid (t-butyl), Fmoc-glutamine(trt), and Fmoc-phenylalanine were all purchased from ChemPep. The N-terminus of each peptide was acetylated, and peptides were cleaved in 95% trifluoroacetic acid (TFA), 2.5% H₂O, and 2.5% triisopropylsilane (Sigma-Aldrich, St. Louis, MO). TFA was mostly evaporated, and peptides were then precipitated twice into cold ethyl ether. Samples were redissolved in water, frozen in liquid nitrogen, and lyophilized. Dried samples were then reconstituted in water and purified by preparative-scale reverse-phase high-performance liquid chromatography (RP-HPLC) using a Waters Xbridge BEH130 Prep C-18 column. The mobile-phase comprised gradients of degassed, deionized water with 0.1% TFA and acetonitrile with 0.1% TFA, at a flow rate of 21 ml/min. Peptide was detected by UV absorbance at 214 nm, and fractions were collected and lyophilized. Molecular weights of the purified peptides were verified by electrospray ionization mass spectroscopy (ESI-MS).

Circular dichroism spectroscopy (CD)

Experimental characterization of the average secondary structure of peptide samples was conducted via circular dichroism spectroscopy on a Jasco 810 circular dichroism spectropolarimeter (Jasco Inc, Easton, MD, USA). Peptides were dissolved in 10 mM potassium phosphate at pH 7.4 with a final peptide concentration of 0.125 mg ml⁻¹. Ionic strength was adjusted for select sample preparation by addition of 500 mM potassium chloride or potassium fluoride stock solutions. Samples were briefly sonicated to aid in the dissolution of the lyophilized peptides, and CD spectra were recorded using a quartz cell with 0.1 cm optical path length. Samples for full wavelength scans at various temperatures were cooled for three minutes at 0 °C prior to the start of the experiment. Scans were recorded from 0 °C to 80 °C, at 10 °C increments, with a step-and-hold heating rate of 1 °C min⁻¹ between temperatures. Samples underwent subsequent cooling to 0 °C at the same increments and cooling rate between isothermal hold steps. For each wavelength scan, the scanning rate was 50 nm min⁻¹, with a response time of 4 s. Wavelengths from 195 nm to 250 nm were recorded at increments of 0.5 nm.

Measurement of peptide unfolding was also conducted by recording the mean-residue ellipticity (MRE) values at 222 nm ($[\Theta]_{\text{MRE},222}$) every 0.5 °C, from 0 °C to 80 °C, while the temperature was increased at a constant rate of 1 °C min⁻¹. Samples were subsequently cooled back to 0 °C at 1 °C min⁻¹ while recording $[\Theta]_{\text{MRE},222}$. In some cases, peptide solutions at higher peptide concentrations (100 μM or 1000 μM) were prepared as above, and then were incubated at 60 °C for two weeks to observe whether slow conformational transition(s) or changes in aggregation state occurred. Full wavelength scans of these incubated samples were performed as described above, but at a single temperature of 60 °C.

Coarse-grained 4bAA model

In order to predict and model the unfolding and self-association of peptides, along with the thermodynamics of the steps involved in those processes, an implicit-solvent CG molecular model was used. It was a modified version of the 4bAA force field proposed by Bereau and Deserno [16] that was extended previously to include long-ranged screened electrostatic interactions [7,20]. Each amino acid is represented as the conjunction of four spherical beads as follows: one for the amide group (N), one for the alpha carbon (C_α), one for the carbonyl group (C') and one for the side chain group (C_β) as shown in Scheme 1. The first three beads correspond to the peptide backbone and are able to interact via steric interactions, bond stretching and bending, and hydrogen bonding [7]. The last bead represents the side chain, with the exception of glycine where no fourth bead is included. The side-chain bead is used to capture the specificity of interactions between side chains, as well as the charge and relative hydrophobicity of each residue. Interactions between beads include local and non-local effects, as follows.

Local interactions correspond to bond distances (2-body interactions), bond angles (3-body interactions), and torsion and improper angles (4-body interactions) due to the planarity of the peptide bond. Local interactions exist only between beads that are covalently bonded to each other. Non-local interactions account for steric repulsions, hydrophobic attractions, hydrogen bonding, and electrostatic interactions that occur between beads that are not covalently linked to each other. Previously, the interaction parameters (other than electrostatics) were parametrized against NMR and crystallographic data in order to capture the secondary and tertiary structures of different polypeptides in their folded state(s) [16]. The electrostatic interactions were

parameterized based on experimental light scattering data to give accurate values of osmotic second virial coefficients for globular proteins as a function of ionic strength [20]. This extended Bereau Deserno (EBD) coarse-grained model treats solvent (water + buffer + salts/solutes) implicitly in order to make the computations tractable for the range of different sequences and solution conditions of interest here and for future work. **Effects of different salts are only captured in a mean-field manner, by accounting for deviations from the Debye-Huckel limiting law for primitive ions in aqueous solution [34].**

The force field for the EBD model is given by the linear combination of each contribution to the interactions:

$$W = \sum_{i<j} u_{ij}^{\text{bond}} + \sum_{i<j<k} u_{ijk}^{\text{angle}} + \sum_{i<j<k<l} (u_{ijkl}^{\text{tors}} + u_{ijkl}^{\text{imp}}) + \sum_{i<j} u_{ij}^{\text{sterics}} + \sum_{i<j} u_{ij}^{\text{hp}} + \sum_{i<j} u_{ij}^{\text{hb}} + \sum_{i<j} u_{ij}^{\text{elec}}$$

where W is the total potential energy (strictly, the potential of mean force) for a given configuration of molecule(s) in the simulation. u_{ij}^{bond} and u_{ijk}^{angle} correspond to the bond-length and bond-angle interactions between two and three contiguously bonded beads, respectively. u_{ijkl}^{tors} and u_{ijkl}^{imp} correspond to the torsion and improper angles interactions due to the backbone constraints. Those terms restrict the possible secondary structures of the peptide through the torsion angles ϕ , ψ and ω (Scheme 1) and the stereoisomer constraints of an amino acid (i.e., L- or D- side chain) [7]. The last four terms in Eq. 1 corresponds to the steric (u_{ij}^{sterics}), hydrophobic (u_{ij}^{hp}), hydrogen bonding (u_{ij}^{hb}), and electrostatic (u_{ij}^{elec}) contributions to the potential energy. The i , j , k and l subindices on the summations indicate that the summations are over all i - j pairs, i - j - k triplets or i - j - k - l quartets of beads in each corresponding case. All the beads are subject to steric repulsions, while only the amide and carbonyl groups are allowed to form hydrogen bonds, and only the side chain beads contribute to hydrophobic and electrostatic interactions. More

details about the CG model and parameters can be found in previous work [7]. As described below, the hydrogen bonding parameter was refined as part of the present work, to more accurately capture experimental unfolding thermodynamics of helical polypeptides.

Molecular dynamics simulations

The conformational stability, intra-peptide and inter-peptide interactions of each sequence were evaluated by performing Replica-Exchange Molecular Dynamics (REMD) simulations at constant volume for a fixed number of peptides, coupled to a Nosé-Hoover thermostat to generate a correct canonical distribution for each replica [35]. REMD is a suitable method to accurately calculate the ensemble-averaged properties at defined temperature intervals, as it helps to prevent the simulation from becoming locked in local minima of the energy landscape at relatively low temperatures. REMD can be coupled with Weighted Histogram Analysis Methods (WHAM) to determine the density of states of the system and then calculate thermodynamic properties and ensemble averaged structural properties over the range of simulated temperatures [7,35,36].

All REMD results presented here were simulated in a cubic box including a single peptide or two peptides with the same sequence. The single-peptide simulations were used to assess changes in the conformational stability of the sequences in the ideal dilute regime (no peptide-peptide interactions) where a box length (L) of 18 nm was used to ensure the box size was larger than the size of the completely extended peptide. 2-peptide simulations were implemented to evaluate the effect of peptide-peptide interactions on the conformational stability and self-assembly behavior of selected peptide sequences. In this case, L was set in order to make the effective or local concentration of the peptide solution equal to 1 mM.

However, the term “concentration” in this context is more properly understood as a measurement of “confinement” for the two peptides within the simulation box. Experimentally, high peptide concentrations cause higher probabilities of inter-peptide interactions, a decrease in accessible volume (“crowding effects”), as well as higher probabilities of three or more peptides interacting simultaneously. The former two are captured reasonably via confinement of the peptides at small box lengths. However, the lattermost effect cannot be captured in simulations unless multiple (more than 2) peptides are simulated simultaneously, and this was not done in the present work in the interest of computational time limitations.

Solution pH was held constant at 7.4, as this is relevant to biotechnology applications and also allows one to treat all Lys (K) and Glu (E) residues as charged (1+ and 1- respectively). Ionic strength was kept constant at 20 mM to avoid complete screening of the electrostatic interactions while maintaining experimentally realistic ionic strength conditions. A set of replicas in REMD were distributed between 220 K and 400 K with the total number of replicas adjusted to each simulated peptide(s) and solution conditions so to assure a replica swap acceptance ratio between 30% and 50%. An integration time step of $\delta_t=0.0035\tau$ (~ 1 fs) was used and swaps between replicas were attempted every 1τ for replica-exchange steps. The initial configuration of each replica was chosen randomly with the peptides in a helical configuration. Each simulation employed two thermal equilibrium periods of $5 \times 10^4 \tau$ each, using standard molecular dynamics and REMD, respectively. A sampling period of $1 \times 10^6 \tau$ was performed using REMD, where the molecular configurations and energy values of each replica were stored every 1τ for subsequent analysis.

Heat capacity values were reconstructed by using WHAM to evaluate the density of states to compute the variance in the energy of the system [7,37]. Mid-point of unfolding

temperatures (T_m) and thermodynamic properties (e.g. enthalpies, entropies and free energies of unfolding) were calculated by using a 2-state transition model fitted to the simulated heat capacity values [38]. Ensemble-averaged helix contents were calculated by computing the number of residues in a helical configuration and then averaging over the weighted ensemble of configurations during the simulation. For a residue i to be defined as helical, it had to comply with the following characteristics: (a) it must form a hydrogen bond with its $i+4th$ or $i-4th$ residue, (b) its torsion angle ϕ lies between -150° and -30° , and (c) its torsion angle ψ lies between -90° and 10° [16]. Other order parameters were also calculated from the simulation: the CONGENEAL score, using a perfect helical peptide as a reference [39]; the radius of gyration (R_g) [7]; the number of side-chain contacts, calculated as the number of side-chain beads with energy magnitude $\geq 0.49 \text{ k}_B\text{T}$ (hydrophobic or hydrophilic attractions); the number of charge-charge contacts; the number of C_α - C_α contacts; the total number of hydrogen bonds and the C_α - C_α mean square distance (RMSD). Given the nature of REMD, and the use of an implicit solvent model, it was not possible to reliably infer kinetics or time scales for transitions such as unfolding from these calculations.

Principal component analysis (PCA) was employed to assess the multivariate nature of protein unfolding and solution behavior. All the computed order parameters were combined and subjected to a PCA treatment: the covariance matrix of the normalized order parameters was calculated for each replica. The eigenvalues and eigenvectors of each covariance matrix were subsequently calculated and all the normalized order parameters were subjected to a vector projection using the eigenvectors obtained from the replica closest to the mid-point of unfolding in order to study the same projection (direction of change) for all the replicas. Finally, a histogram analysis was employed to evaluate correlations between principal components (PCs).

The obtained PCs were organized descendingly by eigenvalues (e.g. PC1's eigenvalue > PC2's eigenvalue) as the magnitude of their corresponding eigenvalues are representative of the amount of information captured by each respective PC. This allows to select only the PCs with highest eigenvalues (first PCs) for further analysis, and it was found that more than 80% of the information was contained in PC1 and PC2 together in all cases. Therefore, a combined PC1-PC2 analysis is the focus in the Results and Discussion section. Inclusion of the other PCs beyond PC1 and PC2 did not alter any of the conclusions drawn below. Additionally, cartoons of snapshot structures were added for T values around the T_m for better readability.

Results and Discussion:

As noted in the Introduction, a challenge for both atomistic and CG molecular models is to accurately produce the thermodynamics of unfolding for polypeptides and proteins [22,23,40]. A series of single-peptide simulations using different model parameters were performed for three previously studied polypeptide sequences, in order to test and possibly refine the original model parameters to assure the original or modified model can capture the unfolding thermodynamics at least semi-quantitatively. The polypeptide sequences were taken from Scholtz et al. [41]. The generic formula $Y(AEAAKA)_nF$ was used and values of $n = 3, 4$ and 5 were considered to assess the effects of chain length on helix stability, as was done experimentally by Scholtz et al. [41]. Table 2 provides short-hand notations and sequences for each of the peptides considered in this work. A solution pH value of 7.4 and 20 mM ionic strength were used to match the reported experimental conditions. Although no T_m values were explicitly reported in the previous work, T_m values were extracted from the data by identifying

the temperature at which the second derivative of the mean residue ellipticity at 222 nm was equal to zero.

In order to tune the prior 4bAA model, only parameters affecting non-local forces were modified so as to maintain the prior structural agreement of folded structures with NMR and crystallographic measurements [16]. The parameters that characterize the strength of hydrophobic attractions (ϵ_{HP}) and hydrogen bonds (ϵ_{HB}) were subject to a simple perturbation analysis to assess which of those exerts the strongest effect on the T_m value (see [20] for an analogous approach). The originally formulated values were perturbed $\pm 10\%$ around the previously reported model (see also, [16] and Supporting Information of [7]). Figure 1A shows an example of the simulation results and analysis for the YAF3 sequence from Scholtz et al., where the constant volume heat capacity (c_v) is plotted as a function of temperature (T) and T_m is identified as the T value at which the heat capacity reaches a maximum. From Figure 1A, it is clear that ϵ_{HB} is the most significant parameter affecting the T_m . This can be anticipated, as unfolding transitions for helical peptides necessarily require breakage of back-bone hydrogen bonds [25,42].

Based on those results, simulations at values of $\epsilon_{\text{HB}} = 4.5, 5.0$ and $5.5 \text{ k}_B\text{T}$ were performed for the YAF3, YAF4 and YAF5 sequences to find the optimum ϵ_{HB} value that allowed experimental and predicted T_m values to align quantitatively. Note that because this is an implicit-solvent model, all HB energy values are inherently relative to water-peptide HB energies. Figure 1B shows the simulated T_m values as a function of the ϵ_{HB} parameter while Figure 1C shows the comparison between experimental and simulated T_m for the same ϵ_{HB} values. A simple linear interpolation was used to find the optimal ϵ_{HB} value that matches both experimental and simulated T_m values. The resulting hydrogen bond strength was $\epsilon_{\text{HB}}=5.09$

$k_B T$, which was used for subsequent simulations. For reference, this is approximately 85% of the value used in previous work [7,16].

Following the tuning of the model, the AAQEAAAQAAAQEAAA (AQE) sequence was originally target based on the utility of similar sequences for pH- and temperature-modulated folding and assembly of related polypeptides. However, this sequence was ultimately not used in the present work as it was found to aggregate spontaneously (data not shown). To promote higher solubility, a lysine residue was added at position 10, resulting in the sequence AAQEAAAQKAAAQEAAA (AQEK). This central lysine also provides amine functionality for common n-hydroxysuccinimide (NHS) reactions for PEG conjugation that are of interest in future work. For additional sequences, phenylalanine residues were substituted for alanine residues at positions 5 and 15 (A5F and A15F mutations) to provide hydrophobicity, and pi-stacking capability, a hallmark of amyloid fibril formation. This substitution yielded the sequence AAQEFAAAQKAAAFQEAAA (FAQEK). A series of variants with different sequence lengths were synthesized using AAAQ repeats with terminal lysines to provide solubility, and a central lysine was used for the two longest sequences. As an additional consideration, glutamic acid residues were eliminated from the sequences to provide a uniform charge. The resulting series of peptide sequences are summarized in Table 1, and were selected to evaluate the capabilities of the updated CG model to capture the effect of (i) selective point mutations for residues with different hydrophobicity and (ii) modifications in the length of the polypeptide. AQEK and FAQEK sequences were selected based on the former, while the AQK18, AQK27 and AQK35 sequences were selected for the latter. Additionally, this particular set of sequences allows evaluation of the sensitivity of the tuned computational approach for thermodynamic properties that compare to those reported from prior work [7,16].

In order to validate the thermodynamic tuning and better understand conformational stability of the proposed sequences, REMD simulations were carried out at pH 7.4 and 20 mM ionic strength for each of the five sequences listed in Table 1 using one peptide in the simulation box for a given sequence.

A natural output from REMD simulations with WHAM is the polypeptide heat capacity (c_v) as a function of temperature (T). Using $c_v(T)$ profiles makes no assumptions regarding a 2-state or multi-state model for the process of unfolding, and allows one to easily characterize T_m and assess the cooperativity of the process via the location of the peak position and sharpness of the transition [7,38,43,44]. One can subsequently analyze configurations from selected temperatures along the $c_v(T)$ profile to deduce the structural changes that occur during thermal unfolding [38].

Illustrative results are given in Figure 2A, where $c_v(T)$ is shown for each of the five peptide sequences. Inspection of Figure 2A shows that each of the five sequences displays reasonably 2-state unfolding behavior, based on the observation of a single, relatively sharp and symmetric peak in each case. Configurations at temperatures significantly lower than T_m correspond to the predominantly folded states, and while those significantly above T_m correspond to unfolded states. This can be visualized from a structural perspective by plotting the average helix content (see Methods) as a function of temperature (Figure 2B), where the transition from folded to unfolded states is essentially a sigmoidal function, as expected for an idealized 2-state unfolding transition [38]. Additionally, the T_m values from the $c_v(T)$ plots correspond to a 50% change in the average helix content from the folded to the unfolded state. For a 2-state transition, T_m is also expected to correspond to the temperature at which the

inflection point occurs in Figure 2B for a given sequence, and this agrees with the standard analysis of CD data [41].

Based simply on considerations of peptide length, an unexpected result is the noticeable difference in T_m that was shown when comparing the AQEK, FAQEK and AQK18 sequences. The AQEK sequence showed a considerably lower T_m than the FAQEK and AQK18 sequences. From inspection of Figures 2A and 2B, it is clear that the length of the peptide sequences considerably affects the thermal stability of the peptide. The T_m increases greatly as the chain length increases, as shown in Figure 2C, and as expected based in prior experimental results [41]. This was initially assessed by evaluating the differences in Gibbs energy (ΔG_{un} , Figure 3A), enthalpy (ΔH_{un} , Figure 3B) and entropy (ΔS_{un} , Figure 3C) between the folded and unfolded configurations as a function of temperature obtained by fitting a 2-state transition model to the obtained $c_v(T)$ plots. Those three thermodynamic properties correlate with the length of the sequence (longer chains resulted in higher values of ΔG_{un} , ΔH_{un} , and ΔS_{un}) so a much higher entropy change upon unfolding occurs as the chain length increases, and this is compensated by a much higher enthalpy change upon unfolding. Overall, this leads to a net increase in the Gibbs energy of unfolding and T_m as chain length increases. Despite having a higher enthalpy of unfolding, the AQEK sequence showed a much lower T_m than the FAQEK and AQK18 sequences (Figure 3B). Figure 3 indicates that these differences in T_m are mainly caused by a considerable difference in the entropy of unfolding. However, the energetic behavior also correlates with the trends in (all but the AQEK) T_m values (Figures 2C and 3B), so a balance between energetic and entropic behavior was considered next.

This was assessed by evaluating the molecular events involved in the unfolding transitions of these sequences. Changes in potential energy and non-local contributions within

the polypeptide chain were evaluated in order to further understand the molecular events that lead to differences in unfolding thermodynamics. It is useful to point out that the local contributions (i.e. average bond lengths and angles, and torsional and improper angles) are the same for all of the sequences, so differences in the simulated thermodynamic properties should arise mainly from non-local contributions (i.e. sterics, hydrogen bonding, side chain hydrophobic interactions and electrostatics). Although the average, total potential energy is used to compute the heat capacity, it does not show any features that conclusively explain significant changes in T_m (Figure 4A). The hydrogen bonding (HB) energy (cf., Figure 4B) reflects a similar trend observed from Figure 2C, suggesting that the increased number of HB interactions as chain length increases is mainly responsible for the increased T_m and the enthalpy of unfolding, in accordance with standard arguments [38,41,45].

However, this does not sufficiently explain the differences between AQEK, FAQEK and AQK18. The first two sequences showed equal H-bonding baselines (Figure 4B) while the third one showed weaker H-bonding energy as expected for a shorter peptide, and that would seem to contradict the results in Figure 2C. In that regard, the contribution from hydrophobic interactions between side chains (cf., Figure 4C) also shows a correlation with changes in T_m . Stronger attractive (non-electrostatic) side-chain interactions lead to higher T_m values, as expected from prior results [46–48]. However, caution is needed before concluding that experimental T_m values should scale with larger hydrophobic interactions between side chains. Stronger hydrophobic interactions of exposed side chains in the folded or unfolded states are expected to lead to stronger polypeptide self-association, which may be unavoidable at finite concentrations needed for experimental measurements of unfolding. The latter typically leads to lower T_m values [6,46,49]. Finally, the electrostatic contributions (Figure 4D) do not show any

correlation with observed T_m values, and the energy values are two orders of magnitude smaller than the other two contributions. Therefore, those contributions are considered effectively negligible in terms of the thermodynamics of the unfolding process of these peptides. Consequently, the results support the view that a balance between the chain length (and number of HB contacts) and side-chain hydrophobicity predominates the conformational stability of these peptides. While the former increases both H-bonding and hydrophobic energies for chemically similar sequences, the latter can alter the position of the T_m by a few degrees without modifications of the length of the chain.

The above points notwithstanding, by analyzing the shape of the curves in each panel in Figure 4, an interesting behavior is notable for the FAQEK and AQK35 sequences in comparison to the other three sequences that could be useful in explaining the deviations from the simple length-dependent scaling in Figure 2. First, all the sequences appeared to follow a 2-state transition (Figures 2A, 2B, 4A, and 4B), which is consistent with the experimental results for those sequences shown and discussed below. Figures 4C and 4D show a different behavior. The hydrophobic contribution shows sigmoidal behavior for the AQEK, AQK18 and AQK27 sequences but it exhibits slightly different behavior for the FAQEK and AQK35 sequences. The FAQEK sequence shows a subtle increase in hydrophobic energy without a defined upper base line as temperature rises, despite large changes and defined base lines in H-bonding energy and average helix content at the same temperatures.

Together, these indicate that the unfolding transition might not be representative of an idealized 2-state transition. That is supported by Figure 4D, where a maximum in the electrostatic contribution is observed for FAQEK above T_m so the sigmoidal behavior is lost. Such a loss is not observed for AQEK despite its similar sequence and identical location of

charged residues. This maxima suggest that FAQEK is subject to a hydrophobic collapse which allows the equally charged residues (E) to approach closer while unfolding. Consequently, the addition of two F residues changes the unfolding events compared to AQEK, and this increases the T_m by considerably decreasing the entropy of unfolding. Similarly, AQK35 shows a maximum in the charge energy, which suggests that the charge-charge distances (K-K distances) decreases as the peptide unfolds. This can only be explained if there is a hydrophobic collapse after unfolding in a similar way as that of FAQEK, which causes K residues to similarly become closer during unfolding. Conversely, AQK18 and AQK27 shows a similar behavior to AQEK as the electrostatic energy decreases during unfolding following a sinusoidal transition, which suggests that the K residues are overall moved further apart during unfolding. Consequently, this analysis suggest the possible existence of additional states during unfolding for the FAQEK and AQK35 while validates the former assumption of an idealized 2-state transition for the AQEK, AQK18 and AQK27. Therefore an additional approach was taken to elucidate the proposed hypotheses.

PCA was performed with the order parameters obtained from the simulations to evaluate the molecular / structural basis for changes in unfolding thermodynamics across the different sequences. In Figure 5, the normalized probability (Π) of observing principal component 1 (PC 1) and principal component 2 (PC 2) is plotted (log scale) in a contour plot as a function of PC 1 and PC 2 for AQEK and FAQEK – i.e., $\log_{10} \Pi$ (PC 1, PC 2) vs PC 1 & PC 2. For 2-state unfolding transitions, only two well-identified and well-populated states or regions should be observed in this type of plot. The results for AQEK and FAQEK show markedly different behaviors in Figure 5, where panels A-C and D-E show, respectively, representative probability surfaces for AQEK and FAQEK, as a function of temperature. The probability surfaces are for T

$\ll T_m$ (panels A and D), $T \sim T_m$ (panels B and E) and $T \gg T_m$ (panels C and F). Inspection of panels A to C shows that AQEK follows a reasonably ideal 2-state transition, as only two main regions are observed. On the other hand, FAQEK shows intermediate states, as four regions are observed in Figure 5E and two regions are observed in Figure 4F.

Structures obtained from the MD simulation show that the intermediate states are represented by a loop formed around the K residue for FAQEK during unfolding, leading to a configuration resembling a small molten globule at temperatures in the vicinity of T_m (Figure 5E and Supporting Information). These configurations are promoted by the lower energy that both F residues allow the peptide to obtain during collapse. This collapse causes both E residues to come closer than in the folded and fully-unfolded, expanded states, which is responsible for the maximum observed in Figure 4D for the red dashed curve. Additionally, two intermediate configurations were observed: where the F residues interact with one another and where F residues interact with any nearby alanine residue. Those intermediate configurations are responsible for the decrease in the entropy of unfolding by not allowing the peptide to fully explore other unfolded configurations. These findings agree with experimentally observed events of protein unfolding in other systems, where molten globules are usually observed before the protein fully unfolds or refolds [46,50]. The inclusion of more hydrophobic residues caused an increase in T_m by limiting the number of extended configurations that the unfolded state could populate, rather than simply lowering the energy of the folded state. While some of the loss of entropy in the unfolded state is compensated by the favorable interactions (lower energy) within the unfolded state, the net result is that the unfolded state(s) are destabilized enough compared to the folded state that higher temperatures are needed to achieve complete unfolding in the simulations for FAQEK, compared to AQEK.

Turning to the series of peptides with common sequence and different lengths, PCA was also carried out for the AQR18, AQR27 and AQR35 simulations as a function of temperature. AQR18 shows a clear 2-state transition similar to that observed for AQEK (Figure 6). Using the T_m of AQEK as a reference, Figure 4 and the discussion above indicate that the higher T_m for the AQR18 results from: (i) decreased enthalpy of unfolding that results from the breaking of fewer hydrogen bonds as well as from higher hydrophobicity in the sequence (Figures 3B, 4B and 4C), (ii) a significantly reduced entropy of unfolding due to the shorter sequence (fewer configurations), and (iii) less extended structures in the unfolded state due to the sequence higher hydrophobicity in comparison with the AQEK. For AQR27, there is a more apparent 2-state transition similar to that observed for AQEK (Figure 7). However, a reasonably well populated intermediate was observed. This arises from the higher likelihood of collapse of the longer chain in the unfolded or partly unfolded state(s), in comparison to the shorter sequences, as well as the addition of a charge residue (K) in the center of the sequence. This also arises from an increased number of stabilizing contacts in (partly) collapsed states for the longer sequence. The partly collapsed configurations resemble those in the unfolding intermediates observed in analysis of the FAQEK behavior, although the intermediates for the AQR27 are less stable due to the weaker hydrophobic interactions (Figures 5E and 7B, and Supporting Information). Similarly, the AQR35 sequence shows a series of intermediate states resulting from chain collapse (Figure 8). These observations are consistent with other experimental and simulation results in which molten globules have been observed during refolding [40,50]. In this context, the AQEK sequence is effectively an outlier in Figure 2, as this peptide does not show such behavior and its unfolding represents an idealized 2-state transition.

In order to experimentally test the simulation results presented and discussed above, the same set of Ala-rich peptides was prepared and evaluated under the same solution conditions as those in the simulations (see Materials and Methods). The helical content and unfolding of each of the peptides were characterized *via* CD spectroscopy, heating the peptide solutions from 0 °C to 80 °C at a concentration of 0.125 mg ml⁻¹ (yielding the following molar concentrations: 71 μM for AQEK, 66 μM for FAQEK, 74, μM for AQK18, 50 μM for AQK27, and 40 μM for AQK35). Representative full wavelength spectra for AQK35 at pH 7.4 are shown in Figure 9A and 9B; the full wavelength spectra for the other sequences are shown in the Supporting Information. At low temperatures, all peptide sequences showed spectra with characteristic α -helical features, with the minima at 208 nm and 222 nm. For sequences AQK18, AQK27, AQK35 a clear isodichroic point was observed, indicating a two-state transition from α -helix to random coil (Supporting Information). Alternatively, an isodichroic point was not clearly observed for the AQEK and FAQEK sequences, signifying the presence of unordered states as the peptides unfold, in partial agreement with the discussion above. The $[\Theta]_{\text{MRE}}$ value at 222 nm at 0° C was used to calculate the fractional helicity using a previously reported method that is based on idealized, long helices (Table S1 in Supporting Information) [51].

Samples were estimated to have average helical contents at 0 °C, relative to an ideal helix, of 12%, 23%, 35%, 40%, and 48% for FAQEK, AQEK, AQK18, AQK27 and AQK35, respectively. While the helical content increased reliably with peptide length in both the simulations (above) and experiments, the experimental helical content values are not equivalent to the simulated average helix contents owing to necessary differences in how the values are determined. Helical content determined experimentally from the CD data is based on comparison to a hypothetical perfect helix, while simulated average helix contents are based on measured

torsion angles and hydrogen bonds (see Materials and Methods) within the simulation. The helicity estimates from CD, however, are in agreement with values reported previously for other short, Ala-rich peptides of similar length [41,45].

The FAQEK sequence showed a lower percent helicity in comparison to AQEK and AQK18. This differs from the simulated peptide model, where the lowest helical content was observed for the AQEK sequence. A previous experimental study reported a similar trend, where inclusion of F residues in a short, Ala-rich peptide was observed to lower the helical content of a peptide compared to identical sequences lacking a phenylalanine residue [52]. This behavior is not observed in the single-peptide simulations, suggesting that inter-peptide interactions can play a role in the stability and unfolding of the FAQEK and other sequences with highly hydrophobic residues. For the other two peptides (AQEK and AQK18), the differences observed in the simulations were borne out experimentally, with AQEK having lower percent helicity than AQK18 despite their similar lengths. As discussed above, this is caused by the stronger hydrophobic contacts within the AQK18 resulting in lower enthalpies and entropy of unfolding. Since the entropy decreases more than the enthalpy, the Gibbs energy of unfolding increases in comparison to the AQEK sequence.

The stability of each of these peptides was also characterized experimentally to assess whether they are prone to aggregation and β -sheet formation. Peptide solutions were heated over a series of temperature values, and absolute intensity values of the $[\Theta]_{\text{MRE}}$ value at 222 nm and 208 nm decreased sigmoidally with increasing temperature, indicating unfolding of the peptide (Figure 9C). $[\Theta]_{\text{MRE},222}$ values were monitored upon heating from 0 °C to 80 °C, as well as upon subsequent cooling back to 0 °C, to analyze the reversibility of unfolding. At pH 7.4, all peptides recovered their original spectrum and $[\Theta]_{\text{MRE},222}$ upon cooling to 0 °C, indicating the

conformational transitions are reasonably reversible on the timescales of the measurements (see Supporting Information). Quantitatively accurate midpoint unfolding temperatures could not be reliably measured, as the pre-transition baseline for each peptide solution was not accessible at temperatures above freezing. This indicated that at least a fraction of the peptides are significantly unfolded at 0 °C, and the values of T_m from these experiments are thus only treated as an “apparent T_m ” values. That notwithstanding, the inflection points (marked with arrows) in Figure 9C show a shift towards higher temperatures with increasing peptide chain length, which is also observed in the simulations. Such an increase in the experimental apparent midpoint unfolding temperature was also observed previously for (AEAAKA)_n sequences, which demonstrated reversible unfolding, consistent with the results from the simulations here [45].

The unfolding curves of AQEK, FAQEK, and AQK18 were similar to the unfolding curves of sequences with shorter chain lengths (14 and 20 residues), in that they exhibited only a portion of the transition region, no flat/linear pre-transition region above 0 °C, and showed a clear post-transition region at higher temperatures. AQK27 and AQK35 also had similar unfolding curves compared to the longer (AEAAKA)_n repeats (26, 32, 38, and 50 residues), with only a portion of the pre-transition region (upper baseline) being observable. Together, the trends obtained from the experimental characterization of AQEK, AQK18, AQK27 and AQK35 agree with the results obtained from the coarse-grained molecular simulations.

However, a conspicuous discrepancy in comparison to the simulated results was apparent for the FAQEK sequence: the experimental unfolding transition is lower than those of any of the other sequences. This discrepancy is likely a result of inter-peptide interactions in the experimentally probed unfolding process, which can cause a decrease in apparent T_m values when the unfolded peptides interact and aggregate [38,49,52,53]. Therefore, preliminary

simulations were performed, where two peptides were present in the simulation box, and the box size was selected to provide an effective peptide concentration of 1 mM. These simulations were used as an initial approach to unveil the discrepancies between observed experimental and single-peptide simulations results, and represent a body of work in progress. Ensemble-average helix content and heat capacity values were calculated as a function of temperature in a manner analogous to the single-peptide simulation (see Materials and Methods) and are presented in Figures 10A and 10B.

Comparison of the results in Figure 10 to those in Figures 2A and 2B clearly show a shift in T_m for the FAQEK in the two-peptide simulation, while the T_m values for the other 4 sequences are not significantly affected by the presence of a second peptide. The results are in better agreement with experimental CD profiles (Figure 9C) and quantitative helical content values discussed above, and support the conclusion that the experimental unfolding behavior of the FAQEK peptides is considerably affected by inter-peptide interactions that cause non-ideal behaviors not considered in the data analysis of DSC and CD experiments.

Finally, an interesting behavior was observed during the 2-peptide simulations, which is reflected in the $c_v(T)$ plots (Figure 10A). Additional peaks and poorly defined unfolded baselines at higher temperatures were obtained for some sequences, in contrast to the single-peptide results (Figure 2A). For all the sequences, it is expected that T_m values from $c_v(T)$ profiles correspond to almost 50% change in the average helix content. However that value was observed below 0 °C for the FAQEK sequence. Consequently, the peak observed at 35 °C in Figure 10A does not correspond to the unfolding and T_m of individual FAQEK molecules. The AQK35 sequence shows a similar peak above 80 °C, and the AQK27 sequence shows a poorly defined baseline at temperatures just above T_m . Inspection of representative configurations from the simulations as a

function of T indicates that these secondary peaks represent the breakage of weak peptide-peptide complexes that had formed during the process of unfolding, and these resemble the initial steps of nucleation of protein aggregation (see Supporting Information) [53].

Finally, samples were made at 1 mM and 0.1 mM peptide at pH 7.4. Samples were incubated at 60 °C for one week and aggregation was qualitatively examined by solution turbidity. Initial experimental results showed aggregates in FAQEK and AQR35 samples, which partially validates this hypothesis and suggests that inter-peptide interaction and self-association of the FAQEK, AQR27 and AQR35 sequences may considerably affect the unfolding characteristics of these peptides (see Supporting Information). This highlights that when modeling idealized unfolding transitions, the models should be complemented with simulations that permit inter-peptide (or inter-protein) interactions. However, given that additional experimental data with higher-level structural resolution are not yet available for the current systems to validate the simulations, it seems unreasonable to extend the interpretation of the multi-peptide effects on unfolding thermodynamics beyond these qualitative conclusions.

Finally, it should be noted that most of the simulations were carried out before the experimental data on the peptides in Table 1 were available, and that the updated EBD coarse-grained model was tuned against a different set of peptides (see Table 2) to allow it to yield predictions that were qualitatively and quantitatively similar to the experimental results for the peptides in Table 1. This suggests that the updated EBD model potentially can be used as an effective means to computationally screen peptide sequences and unfolding thermodynamics for future applications regarding peptide and protein stability, as well as inter-peptide interactions.

Summary

An implicit-solvent 4bAA CG molecular model was successfully tuned to capture unfolding thermodynamics of a series of Ala-rich peptides. This 4bAA model was based on a former model and refined for a set of published peptide sequences. It was further used to provide insight into the unfolding events of a series of new Ala-rich peptides. Initial single-peptide simulations (idealized dilute limit) for AQEK and FAQEK (Table 1) revealed that the inclusion of Phe residues disrupted the idealized 2-state transition exhibited by the AQEK sequence, allowing the formation of stable intermediate states resulting in a decreased entropy of unfolding and higher T_m value. Simulations for the AQK18, AQK27 and AQK35 sequences showed that increases in chain length have a significant impact in the enthalpy of unfolding by increases in hydrogen bonding, which leads to increases in the T_m as the chain length increases as long as the basic chemistry is held constant. Additionally, sequences with charged residues in the middle of the chain showed higher likelihood of unstable intermediate states during unfolding, promoting intermediate collapsed states. CD experiments later showed good agreement between simulated and experimental apparent T_m values for four of the five sequences that were tested. The FAQEK sequence showed deviations that were hypothesized to be attributed to the presence of aggregates within the experimental solutions, suggesting the incorporation of inter-peptide interactions in the analysis of the unfolding behavior. This was corroborated qualitatively by 2-peptide simulations that showed that the interaction between peptides can cause a dramatic change in T_m , and this may be responsible for the low apparent T_m values observed for the FAQEK sequence. This supports the initial hypothesis of aggregation behavior affecting the unfolding thermodynamics of this sequence and encourages the development of computational tools that incorporate both unfolding and aggregation for future work.

Acknowledgements

The National Science Foundation (CHE 1213728) and the National Institutes of Health (R01 EB006006) are gratefully acknowledged for support of this work..

References

- [1] R.M. Murphy, A.M. Tsai, Misbehaving proteins: Protein (mis)folding, aggregation, and stability, Springer, 2006.
- [2] C.J. Roberts, Therapeutic protein aggregation: Mechanisms, design, and control, Trends Biotechnol. 32 (2014) 372–380.
- [3] J.D. Schmit, K. Ghosh, K. Dill, What drives amyloid molecules to assemble into oligomers and fibrils?, Biophys. J. 100 (2011) 450–8.
- [4] A. Ben-Naim, Theoretical aspects of self-assembly of proteins: a Kirkwood-Buff-theory approach., J. Chem. Phys. 138 (2013) 224906.
- [5] G. a Hudalla, T. Sun, J.Z. Gasiorowski, H. Han, Y.F. Tian, A.S. Chong, J.H. Collier, Gradated assembly of multiple proteins into supramolecular nanomaterials., Nat. Mater. 13 (2014) 829–836.
- [6] D.S. Tomar, V. Weber, B.M. Pettitt, D. Asthagiri, Importance of Hydrophilic Hydration and Intramolecular Interactions in the Thermodynamics of Helix-Coil Transition and Helix-Helix Assembly in a Deca-Alanine Peptide., J. Phys. Chem. B. 120 (2016) 69–76.
- [7] B. a. Paik, M. a. Blanco, X. Jia, C.J. Roberts, K.L. Kiick, Aggregation of poly(acrylic acid)-containing elastin-mimetic copolymers, Soft Matter. 11 (2015) 1839–1850.
- [8] E. Sahin, K.L. Kiick, Macromolecule-induced assembly of coiled-coils in alternating

- multiblock polymers., *Biomacromolecules*. 10 (2009) 2740–9.
- [9] A. Top, C. Roberts, K. Kiick, Conformational and aggregation properties of a PEGylated alanine-rich polypeptide, *Biomacromolecules*. (2011) 2184–2192.
- [10] C.J. Roberts, T.K. Das, E. Sahin, Predicting solution aggregation rates for therapeutic proteins: approaches and challenges., *Int. J. Pharm.* 418 (2011) 318–33.
- [11] R. Ghosh, C. Calero-Rubio, A. Saluja, C.J. Roberts, Relating Protein-Protein Interactions and Aggregation Rates From Low to High Concentrations., *J. Pharm. Sci.* 105 (2016) 1086–96.
- [12] N. Agrawal, S. Kumar, X. Wang, Aggregation in protein-based biotherapeutics: Computational studies and tools to identify aggregation-prone regions, *J.* 100 (2011) 5081–5095.
- [13] W. Wang, C.J. Roberts, *Aggregation of Therapeutic Proteins*, John Wiley & Sons, Inc., 2010.
- [14] P.L. Freddolino, C.B. Harrison, Y. Liu, K. Schulten, Challenges in protein folding simulations: Timescale, representation, and analysis., *Nat. Phys.* 6 (2010) 751–758.
- [15] V. Shen, J. Cheung, J. Errington, T. Truskett, Coarse-grained strategy for modeling protein stability in concentrated solutions. II: phase behavior, *Biophys. J.* 90 (2006) 1949–1960.
- [16] T. Berau, M. Deserno, Generic coarse-grained model for protein folding and aggregation, *J. Chem. Phys.* 130 (2009) 235106.
- [17] A. Grünberger, P.K. Lai, M.A. Blanco, C.J. Roberts, Coarse-grained modeling of protein second osmotic virial coefficients: Sterics and short-ranged attractions, *J. Phys. Chem. B.* 117 (2013) 763–770.

- [18] W.G. Noid, Perspective: Coarse-grained models for biomolecular systems., *J. Chem. Phys.* 139 (2013) 090901.
- [19] T. Zhang, P.H. Nguyen, J. Nasica-Labouze, Y. Mu, P. Derreumaux, Folding Atomistic Proteins in Explicit Solvent Using Simulated Tempering., *J. Phys. Chem. B.* 119 (2015) 6941–6951.
- [20] M. a. Blanco, E. Sahin, A.S. Robinson, C.J. Roberts, Coarse-Grained Model for Colloidal Protein Interactions, B 22 , and Protein Cluster Formation, *J. Phys. Chem. B.* 117 (2013) 16013–16028.
- [21] M.S. Shell, Systematic coarse-graining of potential energy landscapes and dynamics in liquids., *J. Chem. Phys.* 137 (2012) 084503.
- [22] T.T. Foley, M.S. Shell, W.G. Noid, The impact of resolution upon entropy and information in coarse-grained models., *J. Chem. Phys.* 143 (2015) 243104.
- [23] M. Cheon, I. Chang, C.K. Hall, Extending the PRIME model for protein aggregation to all 20 amino acids., *Proteins.* 78 (2010) 2950–60.
- [24] N.J.H. Dunn, W.G. Noid, Bottom-up coarse-grained models that accurately describe the structure, pressure, and compressibility of molecular liquids., *J. Chem. Phys.* 143 (2015) 243148.
- [25] S. Chong, S. Ham, Protein Folding Thermodynamics: A New Computational Approach, *J. Phys. Chem. B.* (2014).
- [26] R.S. Farmer, K.L. Kiick, Conformational behavior of chemically reactive alanine-rich repetitive protein polymers., *Biomacromolecules.* 6 (2005) 1531–9.
- [27] R.S. Farmer, L.M. Argust, J.D. Sharp, K.L. Kiick, Conformational Properties of Helical Protein Polymers with Varying Densities of Chemically Reactive Groups.,

- Macromolecules. 39 (2006) 162–170.
- [28] R.S. Farmer, A. Top, L.M. Argust, S. Liu, K.L. Kiick, Evaluation of conformation and association behavior of multivalent alanine-rich polypeptides., *Pharm. Res.* 25 (2008) 700–8.
- [29] A. Top, K.L. Kiick, C.J. Roberts, Modulation of self-association and subsequent fibril formation in an alanine-rich helical polypeptide., *Biomacromolecules.* 9 (2008) 1595–603.
- [30] O. Rathore, D.Y. Sogah, Nanostructure Formation through β -Sheet Self-Assembly in Silk-Based Materials, *Macromolecules.* 34 (2001) 1477–1486.
- [31] M.M. Stevens, S. Allen, M.C. Davies, C.J. Roberts, J.K. Sakata, S.J.B. Tendler, D.A. Tirrell, P.M. Williams, Molecular level investigations of the inter- and intramolecular interactions of pH-responsive artificial triblock proteins., *Biomacromolecules.* 6 (2005) 1266–71.
- [32] W. Shen, K. Zhang, J.A. Kornfield, D.A. Tirrell, Tuning the erosion rate of artificial protein hydrogels through control of network topology., *Nat. Mater.* 5 (2006) 153–8.
- [33] I.W. Hamley, G. Cheng, V. Castelletto, A thermoresponsive hydrogel based on telechelic PEG end-capped with hydrophobic dipeptides., *Macromol. Biosci.* 11 (2011) 1068–78.
- [34] S.I. Sandler, *An Introduction to Applied Statistical Thermodynamics*, John Wiley & Sons, Inc., 2010.
- [35] Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding, *Chem. Phys. Lett.* (1999) 141–151.
- [36] A. Ferrenberg, R. Swendsen, New Monte Carlo technique for studying phase transitions, *Phys. Rev. Lett.* 61 (1988) 2635–2638.
- [37] A. Kadoura, S. Sun, A. Salama, Accelerating Monte Carlo molecular simulations by

- reweighting and reconstructing Markov chains: Extrapolation of canonical ensemble averages and second derivatives to different temperature and density conditions, *J. Comput. Phys.* 270 (2014) 70–85.
- [38] J. Andrews, C. Roberts, Non-native aggregation of α -chymotrypsinogen occurs through nucleation and growth with competing nucleus sizes and negative activation energies, *Biochemistry*. (2007) 7558–7571.
- [39] D.P. Yee, K. a Dill, Families and the structural relatedness among globular proteins., *Protein Sci.* 2 (1993) 884–99.
- [40] T. Zhang, P.H. Nguyen, J. Nasica-Labouze, Y. Mu, P. Derreumaux, Folding Atomistic Proteins in Explicit Solvent Using Simulated Tempering., *J. Phys. Chem. B.* (2015).
- [41] J.M. Scholtz, H.H. Qian, E.J. York, J.M. Stewart, R.L. Baldwin, Parameters of helix-coil transition theory for alanine-based peptides of varying chain lengths in water., *Biopolymers*. 31 (1991) 1463–70.
- [42] A.R. Fersht, V. Daggett, Protein Folding and Unfolding at Atomic Resolution, *Cell*. 108 (2002) 573–582.
- [43] P.L. Privalov, N.N. Khechinashvili, A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study., *J. Mol. Biol.* 86 (1974) 665–84.
- [44] E. Sahin, J.L. Jordan, M.L. Spatara, A. Naranjo, J.A. Costanzo, W.F. Weiss IV, A.S. Robinson, E.J. Fernandez, C.J. Roberts, Computational design and biophysical characterization of aggregation-resistant point mutations for gamma-d crystallin illustrate a balance of conformational stability and intrinsic aggregation propensity, *Biochemistry*. 50 (2011) 628–639.

- [45] J.M. Scholtz, S. Marqusee, R.L. Baldwin, E.J. York, J.M. Stewart, M. Santoro, D.W. Bolen, Calorimetric determination of the enthalpy change for the alpha-helix to coil transition of an alanine peptide in water., *Proc. Natl. Acad. Sci.* 88 (1991) 2854–2858.
- [46] D. Stigter, D.O. Alonso, K. a Dill, Protein stability: electrostatics and compact denatured states., *Proc. Natl. Acad. Sci. U. S. A.* 88 (1991) 4176–80.
- [47] K. a Dill, S.B. Ozkan, M.S. Shell, T.R. Weikl, The protein folding problem., *Annu. Rev. Biophys.* 37 (2008) 289–316.
- [48] E.I. Lin, M.S. Shell, Can peptide folding simulations provide predictive information for aggregation propensity?, *J. Phys. Chem. B.* 114 (2010) 11899–908.
- [49] C.M. Dobson, Principles of protein folding, misfolding and aggregation., *Semin. Cell Dev. Biol.* 15 (2004) 3–16.
- [50] R.R. Goluguri, J.B. Udgaonkar, Rise of the helix from a collapsed globule during the folding of monellin., *Biochemistry.* 54 (2015) 5356–5365.
- [51] R.J. Kennedy, K.-Y. Tsang, D.S. Kemp, Consistent Helicities from CD and Template t/c Data for N-Templated Polyalanines: Progress toward Resolution of the Alanine Helicity Problem, *J. Am. Chem. Soc.* 124 (2002) 934–944.
- [52] A. Chakrabartty, T. Kortemme, R.L. Baldwin, Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions., *Protein Sci.* 3 (1994) 843–852.
- [53] J.M. Andrews, W.F. Weiss IV, C.J. Roberts, Nucleation, growth, and activation energies for seeded and unseeded aggregation of alpha-chymotrypsinogen A, *Biochemistry.* 47 (2008) 2397–2403.

Tables:

Table 1. Synthesized peptide sequences and short-hand notations

Sequence	Short-hand notation	Molecular weight (kDa)
AAQEAAAAQKAAAAQEAAA	AQEK	2.04
AAQEFAAAQKAAAFQEAAA	FAQEK	2.19
K(AAAQ) ₄ K	AQK18	1.95
K(AAAQ) ₃ K(AAAQ) ₃ K	AQK27	2.92
K(AAAQ) ₄ K(AAAQ) ₄ K	AQK35	3.75

Table 2. Peptide sequences use for tuning and short-hand notations

Sequence	Short-hand notation	Molecular weight (kDa)
Y(AEAAKA) ₃ F	YAF3	2.30
Y(AEAAKA) ₄ F	YAF4	2.95
Y(AEAAKA) ₅ F	YAF5	3.59

Scheme 1. Schematic presentation of the 4bAA model based on the Bereau and Deserno model [7,16].

Figure captions:

Figure 1. A: Effect of changing hydrophobic (HP) and hydrogen bonding (HB) parameters for the YAF3 sequence at pH 7.0 and 20 mM ionic strength. **B:** Mid-point of unfolding obtained from the 4bAA model (T_m^{4bAA}) as a function of the hydrogen bonding parameter ϵ_{HB} for the YAF3, YAF4 and YAF5 sequences. Straight lines represent a linear interpolation between the data points. **C:** Comparison between simulated and experimental results [41] of the T_m for three ϵ_{HB} values (4.5 $k_B T$, 5.0 $k_B T$ and 5.5 $k_B T$). The dashed line represents a 1:1 match ($y=x$ function).

Figure 2. A: Heat capacity (c_v) as a function of temperature (T) for single-peptide simulations for different peptide sequences. **B:** Average helix content as a function of T single-peptide simulations. **C:** Mid-point of unfolding temperature T_m , from fitting the simulation results to a 2-state model, as a function of the number of amino acids in the sequence. Lines in panels A and B represent the AQEK (black solid), FAQEK (red dotted), AQK18 (green dashed), AQK27 (blue dashed-dotted) and AQK35 (grey solid) at pH 7.4 and 20 mM ionic strength. The dashed line in panel C represents a linear fit of the four upper data points.

Figure 3. Gibbs energy (panel A), enthalpy (panel B), and entropy (panel C) of unfolding as a function of T obtained from $c_v(T)$ in Figure 1A with a standard 2-state transition model. Different lines represent different peptide sequences: AQEK (black solid), FAQEK (red dotted), AQK18 (green dashed), AQK27 (blue dashed-dotted) and AQK35 (grey solid). All simulated solution conditions are pH 7.4 and 20 mM ionic strength.

Figure 4. Total potential energy (panel A), and its contributions from hydrogen bonding (panel B), hydrophobic attractions (panel C) and electrostatic interactions (panel D) as a function of T obtained from histogram reweighting. The total potential energy includes local interaction energies due to bond fluctuations. Line types are the same as in Figure 2.

Figure 5. Surface plots based on PCA of AQEK at -68 °C (panel A), AQEK at 6 °C (panel B) and AQEK at 117 °C (panel C), compared to FAQEK at -53 °C (panel D), FAQEK at 24 °C (panel E) and FAQEK at 127 °C (panel F) with snapshot structural cartoons around the T_m for better readability. For reference, the T_m values for AQEK and FAQEK are 7 °C and 21 °C respectively. Surface plots represent the normalized histogram for the probability of observing values of PC 1 and PC 2 in a log base 10 scale ($\log_{10} \Pi(\text{PC 1}, \text{PC 2})$). See main text for structural interpretations of the principle components.

Figure 6. PCA of the AQK18 at -69 °C (panel A), 18 °C (panel B) and 112 °C (panel C) with snapshot structural cartoons around the T_m . For reference, the T_m value is 18 °C.

Figure 7. PCA of the AQK27 at -48 °C (panel A), 30 °C (panel B) and 102 °C (panel C) with snapshot structural cartoons around the T_m . For reference, the T_m value is 32 °C.

Figure 8. PCA of the AQK35 at -28 °C (panel A), 42 °C (panel B) and 107 °C (panel C) with snapshot structural cartoons around the T_m . For reference, the T_m value is 45 °C.

Figure 9. Full wavelength spectra of AQK35 during (panel A) heating and subsequent (panel B) cooling at 0.125 mg/ml in 10 mM phosphate buffer (pH 7.4). Peptide samples were heated from 0 °C to 80 °C, and cooled back to 0 °C at 10 °C increments. Full melting curves (panel C) of 0.125 mg/ml peptide solutions in 10 mM phosphate buffer. $[Q]_{MRE}$ value at 222 nm observed while samples were heated from 0 °C to 80 °C. Symbols represent the AQEK (black), FAQEK (red), AQK18 (green), AQK27 (blue) and AQK35 (grey) while colored arrows point to the simulated T_m values from Figure 2C

Figure 10. A: Heat capacity (c_v) as a function of temperature (T) for 2-peptide simulations. **B:** Average helix content as a function of T for 2-peptide simulations. Lines represent the AQEK (black solid), FAQEK (red dotted), AQK18 (green dashed), AQK27 (blue dashed-dotted) and AQK35 (grey solid) at pH 7.4 and 20 mM ionic strength.

Figure 1. See list of Figure Captions

Figure 2. See list of Figure Captions

Figure 3. See list of Figure Captions

Figure 4. See list of Figure Captions

Figure 5. See list of Figure Captions

Figure 6. See list of Figure Captions

Figure 7. See list of Figure Captions

Figure 8. See list of Figure Captions

Figure 9. See list of Figure Captions

Figure 10. See list of Figure Captions