Unfolded states under folding conditions accommodate sequence-specific conformational preferences with random coil-like dimensions

Ivan Peran[a,1,2], Alex S. Holehouse[b,1], Isaac S. Carrico[a], Rohit V. Pappu[b,3], Osman Bilsel[c,3], and Daniel P. Raleigh[a,d,3]

[a]Department of Chemistry, Stony Brook University, Stony Brook, NY 11794-3400; [b]Department of Biomedical Engineering, Center for Science and Engineering of Living Systems, Washington University in St. Louis, St. Louis, MO 63130; [c]Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605; and [d]Institute of Structural and Molecular Biology, University College London, London WC1E 6BT1, United Kingdom

Edited by William A. Eaton, National Institutes of Health, Bethesda, MD, and approved May 3, 2019 (received for review October 22, 2018)

Proteins are marginally stable molecules that fluctuate between folded and unfolded states. Here, we provide a high-resolution description of unfolded states under refolding conditions for the N-terminal domain of the L9 protein (NTL9). We use a combination of time-resolved Förster resonance energy transfer (FRET) based on multiple pairs of minimally perturbing labels, time-resolved small-angle X-ray scattering (SAXS), all-atom simulations, and polymer theory. Upon dilution from high denaturant, the unfolded state undergoes rapid contraction. Although this contraction occurs before the folding transition, the unfolded state remains considerably more expanded than the folded state and accommodates a range of local and nonlocal contacts, including secondary structures and native and nonnative interactions. Paradoxically, despite discernible sequence-specific conformational preferences, the ensemble-averaged properties of unfolded states are consistent with those of canonical random coils, namely polymers in indifferent (theta) solvents. These findings are concordant with theoretical predictions based on coarse-grained models and inferences drawn from single-molecule experiments regarding the sequence-specific scaling behavior of unfolded proteins under folding conditions.

protein folding | unfolded state | FRET | compaction transition

The mechanisms of protein folding remain a topic of intense interest (1–4). Recently, there has been substantial progress in characterizing folding pathways using a combination of novel experimental approaches and improved computational methods (5–7). However, the initial stages of protein folding, specifically the nature of unfolded states under folding conditions, remain a source of controversy, even for simple single-domain globular proteins that undergo an apparent two-state folding transition (8–12). There is convergence regarding the nature of unfolded states under highly denaturing conditions. These states are well described by statistics of self-avoiding walks that are congruent with polymers in good solvents. These highly denatured states are expanded with reduced local and long-range interactions compared with the native state (13, 14). In contrast, the nature of the unfolded state under folding conditions remains enigmatic and the topic of intense debate. A convergent description of unfolded states under folding conditions would be of direct relevance to understanding the mechanisms of protein folding and interactions involving unfolded proteins in cellular settings.

The folded states of globular proteins are typically compact, and their global dimensions are on average well described as polymers in poor solvents (15). Upon dilution from high concentrations of denaturant into native conditions, a foldable yet unfolded and expanded protein must undergo a collapse transition to progress to its compact and folded (native) state (16). Rapid contraction or collapse upon dilution has been observed in refolding experiments for some, but not all proteins (8, 9, 11, 17). Determining the position of the collapse transition along the folding reaction coordinate and characterizing the conformational properties of unfolded ensembles before and following this collapse are essential steps toward deciphering folding mechanisms and for developing a clear understanding of how the amino acid sequence determines the solution behavior of proteins (16).

Proteins are marginally stable molecules. Accordingly, folded and unfolded states are accessible via spontaneous fluctuations even under folding conditions. Fluctuations into and out of unfolded states under folding conditions have been demonstrated using single-molecule spectroscopies (12, 18–22). Further, in physiologically relevant milieus, the interactions within and between unfolded proteins are of functional relevance. These interactions are likely to be influenced by the amplitudes of conformational fluctuations that define unfolded ensembles under folding conditions. As an example, it is now well established that the properties of unfolded states influence the tendencies of proteins to aggregate, and this has important implications for human disease and biotechnology (23, 24). While our focus here is on the unfolded states

Significance

The tools of structural biology afford high-resolution descriptions of folded states of proteins. However, an atomic-level description of unfolded states under folding conditions has remained elusive. Challenges arise from the pronounced conformational heterogeneity and the very low population of unfolded states under folding conditions. We have combined a series of time-resolved biophysical experiments with all-atom simulations and polymer theory to obtain a high-resolution description of unfolded ensembles under folding conditions. These unfolded states are characterized by discernible sequence-specific conformational preferences. These preferences are averaged over by conformational fluctuations, giving rise to ensemble-averaged properties that are consistent with those of canonical random coils. Our findings are relevant to understanding functional and pathological interactions involving unfolded forms of proteins.

Author contributions: I.P., A.S.H., I.S.C., R.V.P., O.B., and D.P.R. designed research; I.P., A.S.H., and O.B. performed research; I.P., A.S.H., I.S.C., and O.B. contributed new reagents/analytic tools; I.P., A.S.H., I.S.C., and O.B. analyzed data; and I.P., A.S.H., R.V.P., O.B., and D.P.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]I.P. and A.S.H. contributed equally to this work.

[2]Present address: Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, TN 38105.

[3]To whom correspondence may be addressed. Email: pappu@wustl.edu, osman.bilsel@umassmed.edu, or d.raleigh@ucl.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1818206116/-/DCSupplemental.

Published online June 5, 2019.

of foldable proteins, numerous functional proteins are characterized by significant conformational heterogeneity even under native conditions. For these intrinsically disordered proteins (IDPs), a better understanding of the unfolded states under native conditions of intrinsically foldable proteins could provide a set of complementary insights that have the potential of improving our understanding of sequence-to-conformation relationships in systems defined by conformational heterogeneity (25–28).

Under native conditions, the free energy balance strongly favors the folded state. Accordingly, unfolded states are accessible via spontaneous albeit low-likelihood fluctuations (29). Characterization of the resultant low-likelihood states requires the use of methods with high temporal resolution that can "catch" short-lived transitions into and out of the unfolded state. The combination of rapid mixing techniques and spectroscopic measurements has allowed the direct interrogation of unfolded states under near native conditions in the absence of strongly destabilizing mutations (30–32). Small-angle X-ray scattering (SAXS) integrated with stopped-flow or microfluidic mixing has also been used to study the early stages of protein folding upon dilution out of high concentrations of denaturant. SAXS experiments have often, but not always, failed to detect collapse before the folding transition for proteins of <150 residues (9, 11, 33, 34). Förster resonance energy transfer (FRET) experiments, also in combination with stopped-flow or microfluidic techniques, suggest that contraction of the unfolded state can occur before the folding transition (8, 35–38). Single-molecule fluorescence experiments provide the clearest evidence for some proteins undergoing continuous contraction of their unfolded states as a function of decreasing denaturant concentration (19–22, 39–41).

Theoretical studies predict that the nature of the collapse transition is governed by the ratio of the two- to three-body interaction coefficients (42). Below a critical value of this ratio, the collapse transition is abrupt and pseudo-first order. Conversely, above a critical value for the ratio of the two- to three-body interaction coefficients, the collapse transition is continuous and will be governed by sequence-specific interactions (3, 29, 43). While there has been a growing consensus regarding the nature of the unfolded state under native conditions, some experiments have yielded seemingly conflicting views regarding chain compaction. This may reflect the fact that different proteins behave differently, based on the sequence-specific ratio of two- to three-body interaction coefficients. There may also be a contribution from the inherent limitations of each method. SAXS provides a label-free method to obtain insights regarding global dimensions such as R_g, and with high enough signal-to-noise can also provide information about overall shapes and scaling behavior (44–47). However, SAXS also has its limitations: It offers limited structural resolution, requires high protein concentrations, is more sensitive to expanded conformations as a direct consequence of the nature of the averaging involved, and can be insensitive to the presence of low-likelihood, long-range contacts and local structure (48, 49). These limitations are of particular importance when characterizing unfolded proteins under low concentrations of denaturant in equilibrium experiments. Under these conditions the fraction of unfolded proteins is very small and the ensemble averaging inherent to SAXS will prevent the discrimination of subpopulations. An alternative is to perform time-resolved ensemble experiments where the folding time is much longer than the dead time of the experiment—a strategy we pursue in this work.

Single-molecule experiments, especially single-molecule FRET, afford the important advantage of being able to directly access specific pairwise distance distributions—a feature of single-molecule detection (18, 50). This provides a direct way to monitor the evolution of populations of distinct substates in equilibrium experiments as a function of denaturant concentration. In the past, extraction of pairwise distances and global properties from FRET data relied on simple models to translate transfer efficiencies into distances (51–56). Recent advanced approaches have replaced simplified models with sophisticated methods that are aided by atomistic computer simulations and Bayesian methodologies (37, 38, 41, 53, 54). If multiple pairwise interresidue distances can be probed (19, 54), albeit in different constructs, then the discrepant inferences due to the decoupling of end-to-end distances and R_g values that arise due to the heteropolymeric, finite-size nature and the significant contributions from three-body interactions (53, 57) can be overcome to generate coherent insights regarding the dimensions of unfolded polypeptides in the absence of denaturants—a feature that is unique to single-molecule experiments (37, 38, 50, 58). Single-molecule FRET studies do rely on the use of bright dyes, which are invariably large and have large (58) Förster radii (R_0). Appending large dyes connected by flexible linkers might perturb the system and may facilitate dye–protein interactions, although recent work suggests that the effects of dyes are not generic and their sequence-specific interactions can be quantified and accounted for either by changing dyes or by accounting for their effects in data analysis (37, 53, 57–60).

To obtain a high-resolution description of the unfolded state under native conditions we pursued a different approach that combined minimally invasive, time-resolved FRET with time-resolved SAXS, all-atom simulations, and analyses guided by polymer theory. The N-terminal domain of the ribosomal protein L9 (NTL9) was used as a model system, as it shows well-defined two-state folding behavior and has been extensively characterized in the context of protein folding (7, 48, 61–63). NTL9 is 56 residues long and is one of the simplest examples of a common α–β fold (Fig. 1*A*). The domain has been shown to contain residual structure in the unfolded state populated under native conditions (62, 63). We monitored chain contraction in real time using a sensitive and minimally perturbing FRET method that exploits *p*-cyanophenylalanine (F_{CN}) and Trp residues as FRET pairs. F_{CN} is the cyano analog of Tyr and it represents a minimal perturbation to the original sequence, thus making it unlike large dyes used in single-molecule experiments. F_{CN} acts as the donor to Trp with an R_0 of 16 Å (64). The residue can be incorporated into proteins using solid-phase peptide synthesis or recombinantly using the 21st pair technology of Schultz and Mehl (65, 66). F_{CN} fluorescence can be excited selectively in the presence of Trp and Tyr and the fluorescence decay of F_{CN} is single exponential, facilitating the analysis of time-resolved studies.

We used time-resolved FRET to measure multiple pairwise distance distributions for the unfolded state populated under strongly denaturing conditions [10 molar (M) urea] and under native conditions (1 M urea). Continuous-flow methods interfaced with time-resolved detection were used to measure pairwise distance distributions for the unfolded protein in 1 M urea, after rapid dilution out of high denaturant. The FRET studies were complemented by continuous-flow SAXS measurements by leveraging the fact that the folding time for NTL9 is on the order of 2.4 ms, while the dead times for FRET and SAXS measurement are 40 μs and 200 μs, respectively. Finally, a globally consistent ensemble of conformations was generated using all-atom simulations. The data reveal that for NTL9, modest contraction occurs rapidly, on a timescale that is faster than that of folding. The global ensemble-averaged dimensions of the chain are consistent with those of a finite-sized Flory random coil, i.e., a flexible polymer in an indifferent or theta solvent. Despite this statistical feature, the ensembles accommodate long-range interactions and fluctuations into and out of native and nonnative elements of structure. These structural preferences are more persistent than the corresponding contacts sampled in high concentrations of urea (48). This study highlights the power of minimally perturbing fluorescence probes for following rapid conformational changes and the importance of combining multiple pair positions as well as methods to construct accurate descriptions of conformational properties of unfolded ensembles.

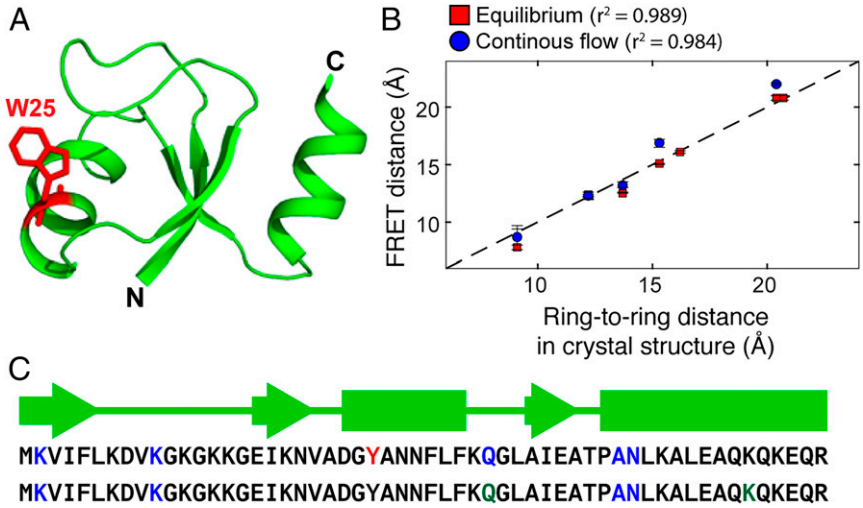

Fig. 1. The structure of NTL9 and the accuracy of distance calculations in the folded state. (*A*) Ribbon diagram of NTL9 native state (PDB code: 2HBB) showing the location of Trp-25. The N terminus is labeled. (*B*) The FRET-derived distances for the native state of NTL9 compared with the ring-to-ring distances taken from the crystal structure, demonstrating the precision with which distances can be determined by Phe_{CN}-Trp FRET. Error bars were determined by rigorous error analysis (*SI Appendix*). Both continuous-flow and equilibrium data are reported. (*C*) Primary sequence of NTL9. (*C*, *Bottom*, first row) The location of Tyr-25, which was mutated to Trp for FRET studies, is shown in red and the location of the five F_{CN} substitutions which are paired with Trp-25 is colored blue. (*C*, *Bottom*, second row) Residues 2 and 33 as well as residues 2 and 51 were used to prepare additional F_{CN}–Trp pairs and the position of these Trp residues is colored green. (*C*, *Top*) The schematic diagram above the sequence illustrates the secondary structure.

Results

The folding time constant of wild-type NTL9 in 1 M urea, pH 5.5, is 2.4 ms, allowing much of the folding process to be accessed using continuous-flow methods. To obtain a high-resolution description of the unfolded state, we introduced a series of point mutations that allowed us to obtain pairwise distance distributions using data from FRET measurements.

Mutations to Introduce FRET Pairs Are Minimally Perturbing to Stability and Folding Rates. A set of seven variants of NTL9 containing F_{CN} and Trp were prepared using 21st pair technology. The native Tyr at position 25 was replaced by Trp and F_{CN} sites were selected to probe a range of through-space distances and to ensure that FRET pairs span a range of sequence separations (Fig. 1*C*). The Trp and F_{CN} residues are on the surface of the protein. The sequence separation between donor and acceptor varies from 8 to 49 residues, while the distance between the sites (C_β–C_β) varies from 8 Å to 21 Å in the native state. All seven mutants were well folded, displayed sigmoidal unfolding profiles as a function of denaturant, and showed similar circular dichroism (CD) profiles to that of the wild-type protein (*SI Appendix*, Figs. S1 and S2 and Tables S1–S4). The folding rate of each mutant was comparable to the wild-type rate (*SI Appendix*, Table S5). The distances derived from equilibrium FRET measurements of the folded protein under native conditions show nearly perfect agreement with the expected distances derived from the crystal structure (Fig. 1*B* and *SI Appendix*, Tables S6–S8), demonstrating that the dyes do not perturb the tertiary structure of the folded state of NTL9. The analysis also shows that precise and accurate distance information can be derived from the FRET studies. These data confirm that the dyes are minimally perturbing and are useful as probes for the evolution of through-space distances.

The Unfolded State Is Expanded in 10 M Urea. We first characterized the equilibrium folded- and denatured-state ensembles under 1 M urea and 10 M urea, respectively. This helps establish baselines for comparisons to data collected in continuous-flow mode. Measurements for the folded state were made in 1 M urea since this corresponds to the final conditions in the continuous-flow mixing studies and represents a facsimile of native conditions inasmuch as the protein population is greater than 99% folded (*SI Appendix*, Fig. S1). The fluorescence lifetime data were fitted to extract distance distributions between donor and acceptor in the folded state and in the high urea unfolded state (Fig. 2 *A* and *B* and *SI Appendix*, Fig. S3). Interresidue distance distributions in the folded state were fitted to a Gaussian distribution. These distributions are, in general, narrow (Fig. 3 and *SI Appendix*, Table S6) and they agree well with the distances expected from the crystal structure (Fig. 1*B*). The highly denatured state ensemble was fitted separately to both a Gaussian distribution and a wormlike chain (WLC) model with diffusion, with both methods yielding nearly identical results (*SI Appendix*, Tables S9 and S10). The distance distributions for the denatured state ensemble in 10 M urea are much broader and show a monotonic increase in mean distance vs. sequence separation. This is to be expected for an expanded chain in high denaturant (Fig. 4).

The Unfolded-State Ensemble Under Folding Conditions Is Contracted Compared with the Denatured State in 10 M Urea. We performed fluorescence lifetime measurements in combination with continuous-flow mixing to enable the collection of fluorescence decays as a function of refolding time (Fig. 2 *C* and *D* and *SI Appendix*, Fig. S4). We observed chain contraction within the dead time of the instrument, indicating that contraction occurs on a timescale that is considerably faster than that of folding. Singular-value decomposition (SVD) analysis yields a maximum of two components, the amplitudes of which can be fitted to a single exponential model. This is consistent with two-state folding (*SI Appendix*, Fig. S3) and it allowed us to perform a global analysis using a two-state model. A Gaussian distribution was used to describe the folded state and either a WLC model or a Gaussian distribution was used for the unfolded state. Both models indicate contraction of the unfolded state upon dilution to 1 M urea. Most notably, the two pairs at greatest sequence separation (F_{CN}2-Trp25 and F_{CN}2-Trp33) suggest a 40% contraction of the unfolded state in 1 M urea relative to the 10 M urea unfolded state. The degrees of contraction suggested by the other FRET pairs range from 12% to 31% (*SI Appendix*, Tables S11–S13). Since highly denatured proteins are well described by the physics of polymers in good solvents, we asked whether similar polymer models would be able to describe the unfolded state under native conditions.

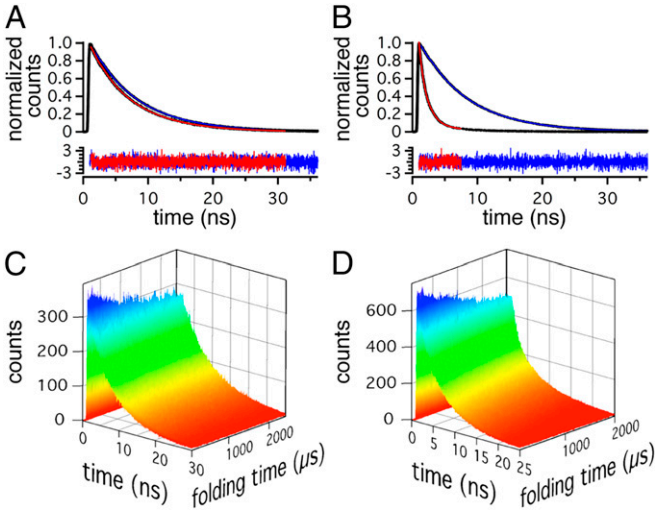

Fig. 2. Representative fluorescence lifetime measurements. Data are shown for the F_{CN}10–W25 pair. For *A* and *B*, fits to donor-only curves are in blue, and fits to donor plus acceptor are in red. Residuals are shown below. (*A*) Equilibrium data in 10 M urea. (*B*) Equilibrium data in 1 M urea. In *C* and *D*, continuous-flow fluorescence decays as a function of folding time are shown. Decays were fitted globally to a two-population model. (*C*) Donor only. (*D*) Donor/acceptor. F_{CN} was preferentially excited at 240 nm and fluorescence was detected by utilizing a Semrock 292/27 bandpass filter, which has a bandwidth of 27 nm centered at 292 nm. Donor-only decays (*A* and *B*, blue) were fitted to a double-exponential model reconvoluted with the instrument response function. The donor/acceptor decay in 10 M urea (*A*) was fitted using a wormlike chain model with a parameter for diffusion. The donor/acceptor decay in 1 M urea (*B*) was fitted to a Gaussian distribution. Donor/acceptor decays from continuous-flow experiments were fitted globally using a Gaussian distribution for the folded state and a wormlike chain model, with a parameter for diffusion between donor and acceptor during the fluorescence lifetime, for the unfolded state.

Fig. 4 describes the relationship between sequence separation and spatial separation for the protein in 1 M and 10 M urea, respectively. In 10 M urea the pairwise-distance profiles are largely consistent with that of a polymer in a good solvent, with a monotonic increase in spatial separation with increasing sequence separation. In contrast, in 1 M urea the distances show a nonmonotonic increase with sequence separation. In fact, we observed a stronger correlation between the distances from the 1-M unfolded state and the folded state than between the unfolded states in 1 M and 10 M urea (Pearson $r = 0.73$ vs. 0.62). Importantly, the nonmonotonic changes in intermolecular distances with sequence separation would, on the surface, seem to be inconsistent with the fractal behavior that is expected of homopolymers in good or indifferent (theta) solvents.

Continuous-flow SAXS data were also collected for wild-type NTL9. A major challenge with these experiments originates from the small dimensions of the protein and the presence of urea, meaning the contribution of the protein to the total scattering is low. The R_g value of NTL9 in 10 M urea was determined from equilibrium experiments to be 23.5 ± 0.7 Å, which agrees with prior equilibrium SAXS studies of the urea-unfolded state (48). The measured R_g value for the folded state in 1 M urea is 12.8 ± 0.2 Å. For comparison, the R_g for the unfolded state in 1 M urea is 19.1 ± 0.9 Å (Fig. 5). This is indicative of ∼20% contraction upon dilution out of high denaturant. The Kratky plot for the folded state shows a peak that is characteristic of compact globules (*SI Appendix*, Fig. S5). Conversely, a monotonic increase with increasing scattering angle is seen for the unfolded state in 10 M urea, indicating a highly expanded chain (*SI Appendix*, Fig. S5). Similar behavior has been observed for other globular proteins in high concentrations of urea and for IDPs with high proline and/or charge contents (67–69). The Kratky plot for the unfolded protein in 1 M urea is very different, and a plateau is observed at increasing q values, indicating a fundamentally different ensemble from the highly denatured state ensemble.

All-Atom Simulations Show That SAXS and FRET Measurements Yield Mutually Consistent Results. At first glance, the time-resolved FRET and SAXS results for NTL9 in 1 M urea might appear to point toward conflicting views of the unfolded state under folding conditions. The FRET data may be interpreted to imply a substantial collapse and/or the presence of interactions that cause significant deviations of intrachain distances from the expected behavior for polymers in theta or good solvents. In contrast, SAXS results demonstrate modest contraction, and given that the R_g value is considerably higher than that of the folded protein, the data are inconsistent with a sharp collapse of the unfolded states. To make sense of the totality of the data, we used all-atom Metropolis Monte Carlo simulations with the ABSINTH implicit solvent model to generate a series of unfolded-state ensembles. We used these simulations to assess the mutual compatibility of results from FRET vs. SAXS experiments. The atomistic descriptions used in these simulations allow us to preserve the sequence-specific interactions that give rise to heterogeneous contact patterns (below) that are easily glossed over if one were to use coarse-grained, single bead per residue descriptions. Importantly, these simulations capture the sidechain-specific backbone conformational preferences and they enable the explicit inclusion of the interplay between two- and three-body interactions, which are nonexistent in preparameterized coarse-grained models.

We generated a set of unfolded-state ensembles for NTL9 at different simulation temperatures. We previously used this approach to show that the ensemble generated at 390 K serves as a good proxy for the solution behavior of NTL9 in 8 M urea (48).

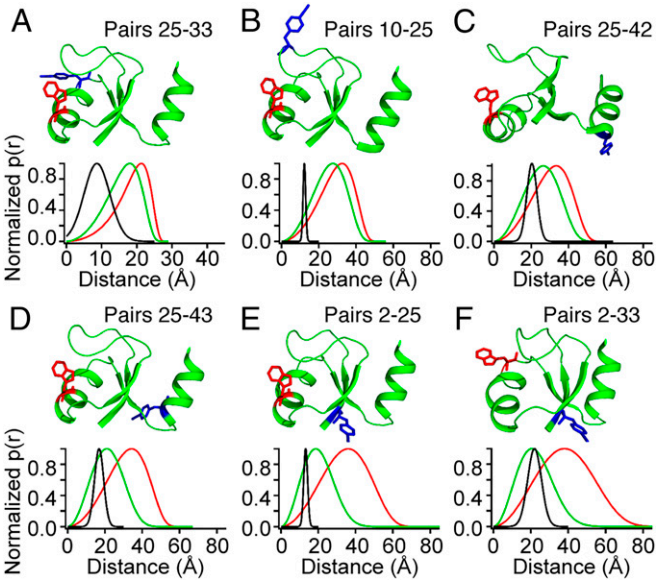

Fig. 3. (*A–F*) FRET provides evidence for compaction. Ribbon diagrams illustrating the location of the FRET pairs are shown together with the distance distributions. Red, unfolded state in 10 M urea; green, unfolded state in 1 M urea; black, folded state in 1 M urea. The folded and unfolded distributions in 1 M urea were extracted from the global fit to the FRET data. The unfolded-state distance distribution was modeled using a wormlike chain with a parameter for diffusion between donor and acceptor during the fluorescence lifetime, while a Gaussian distribution was used to model the folded-state distribution.

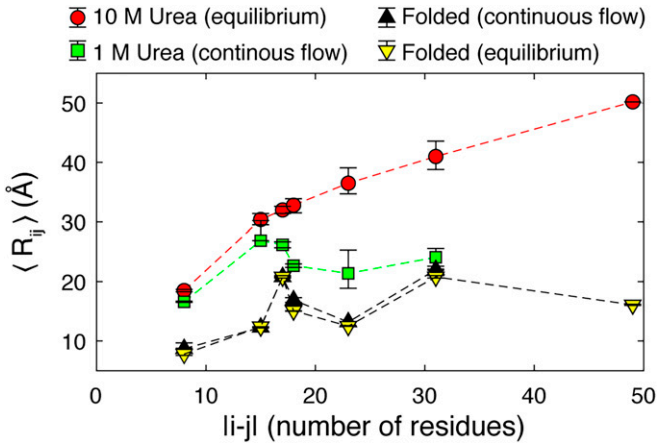

Fig. 4. The 1-M unfolded state is more compact than the 10-M unfolded state. The unfolded state is expanded in high denaturant but is more compact in low denaturant, as quantified by average through-space distances between pairs of residues as a function of sequence separation. The folded state and the unfolded ensemble in 1 M urea both show a nonmonotonic relationship between sequence separation and spatial separation, while the 10-M urea ensemble shows a monotonic increase in distance with sequence separation. See *SI Appendix*, Fig. S7C for the same data which include standard polymer models.

Accordingly, we sought to identify an ensemble where the derived C_β–C_β pair distance distributions taken from the set of donor/acceptor pairs used for FRET matched the experimentally obtained distance distributions. While several ensembles yielded pair distance distributions that were qualitatively similar to the FRET distance distributions, none were quantitatively identical. To alleviate this issue we used a χ^2 minimization and entropy maximization approach (COPER) to reweight each ensemble to match the FRET data for the unfolded state in 1 M urea (70). We then selected the reweighted ensemble that maximized the ensemble entropy while generating the best agreement with the FRET data (see *SI Appendix* for further details). Reweighting was performed based on C_β–C_β distances extracted from the simulated ensembles and distances extracted from analysis of the FRET experiments.

The ensemble that best fitted the data while undergoing minimal perturbations was generated by reweighting the unfolded ensemble generated at 375 K (Fig. 6*A* and *SI Appendix*, Fig. S6). An independent assessment of the accuracy of this ensemble comes from the observation that the mean R_g obtained from this reweighted ensemble (referred to hereafter as the 1 M urea unfolded-state ensemble) is 18.9 Å, which agrees very well with the value for R_g derived from the SAXS data (19.1 Å, Fig. 6*B*). Furthermore, a formal comparison between the scattering profile from the simulation of the 1 M urea unfolded-state ensemble and that from the experimentally measured unfolded state in 1 M urea shows good agreement (*SI Appendix*, Fig. S5). It is worth emphasizing that the SAXS data were not used to obtain the optimized unfolded-state ensembles, yet reweighting improves the χ^2 of the 375-K ensemble with respect to the full scattering curve, formally demonstrating that reweighting using data from our FRET measurements enhances agreement between simulations and SAXS. Our results indicate that inferences based on FRET and SAXS data do in fact yield mutually consistent descriptions for the unfolded-state ensembles.

The Unfolded-State Ensemble Under Folding Conditions Contains Extensive Native and Nonnative Interactions. The computationally derived atomic-level description of the ensemble for the unfolded state in 1 M urea correctly reproduces the FRET and SAXS results, suggesting that it represents a reasonable, high-resolution description of the unfolded ensemble under folding conditions for NTL9. Accordingly, we examined additional features within the ensemble to gain insights regarding conformational features of the unfolded state under folding conditions. The ensemble is characterized by the presence of native and nonnative contacts, as shown in Fig. 7 *A* and *E*. Pairwise contacts primarily involve hydrophobic residues (Fig. 7*C*), with a notable local cluster formed by aromatic and hydrophobic residues around residue 30. The ensemble also accommodates the formation of secondary structure (Fig. 7*E*) and long-range interactions that have finite likelihoods associated with them (Fig. 7*B*). These features have higher probabilities of being observed compared with the ensemble generated to match data for NTL9 in 8 M urea (Fig. 7*B*). Unfolded-state ensembles under refolding conditions show a significant increase in the extent of native and nonnative structural preferences while remaining expanded (Fig. 7*F*). Taken together, these results indicate that the unfolded state under refolding conditions can accommodate pronounced sequence-specific native and nonnative conformational and structural preferences that are enhanced compared with the highly denatured-state ensembles.

Ensemble-Averaged Conformational Properties of Unfolded-State Ensembles Under Native Conditions Are Concordant with Random Coils in Theta Solvents. For flexible homopolymers, a scaling exponent ν quantifies the length scale for correlations among fluctuations between pairs of interresidue distances. The value of ν depends on solution conditions. It quantifies the interplay between intrachain and chain–solvent interactions and is governed by the correlation length for conformational fluctuations and the distance to the theta point. Given the connection between interresidue distances and R_g, it follows that the R_g value and degree of polymerization (chain length) are related by a simple scaling relationship of the form

$$\sqrt{\left\langle R_g^2 \right\rangle} = R_0 N^{\nu}. \qquad [1]$$

Here, $\langle R_g^2 \rangle$ is the mean-squared radius of gyration, N is the number of residues in the protein, and R_0 is a prefactor that is determined

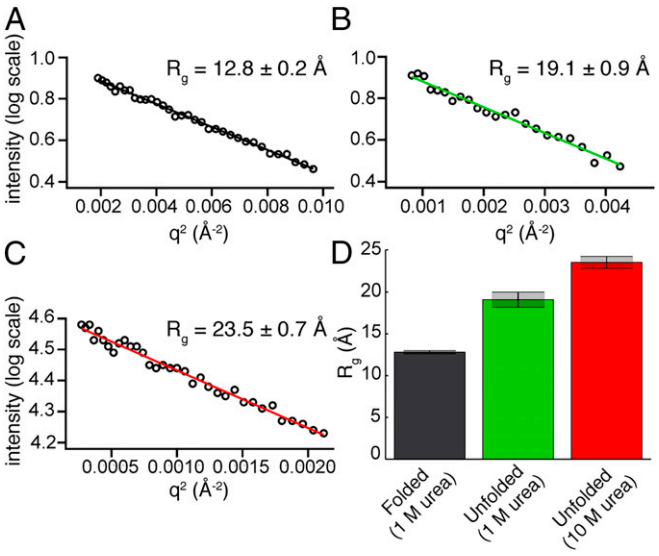

Fig. 5. Guinier analysis of SAXS data. (*A*) Continuous-flow data for the native state in 1 M urea. (*B*) Continuous-flow data for the unfolded state in 1 M urea. (*C*) Equilibrium data for the unfolded state in 10 M urea. Data were also analyzed using an independent empirical molecular form factor, which yielded near identical results (*SI Appendix*, Table S15). (*D*) Comparison of average radii of gyration across the folded and unfolded states. Error bars are those calculated from the Guinier fits.

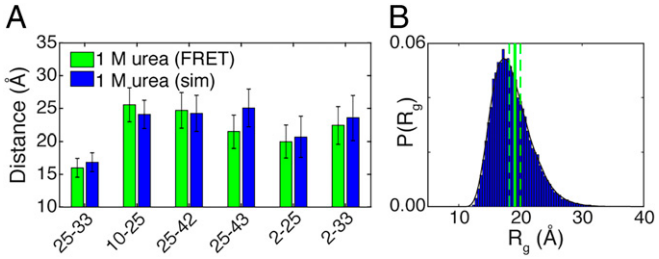

Fig. 6. Comparison between experimental results from the 1-M urea unfolded state and the 375-K reweighted ensemble (unfolded-state ensemble in 1 M urea). (*A*) Intrachain distances obtained from FRET experiments (green) compared with the C_β–C_β distance extracted from the unfolded ensemble (blue). The error bars reflect the SDs of the underlying distribution associated with each measurement. (*B*) Radius of gyration obtained from SAXS with error bars compared with the distribution of radii of gyration obtained from simulation. The simulation data in both *A* and *B* were extracted from the same ensemble, demonstrating that FRET, SAXS, and simulation are mutually compatible.

by the average strength of solvent-mediated interresidue interactions, the thickness of the chain, and the average size of an individual residue. Note the R_0 in this context is not the Förster radius.

In a poor solvent, $\nu = 1/3$ with chain–chain interactions being preferred over chain–solvent interactions, leading to compact and globular ensembles (71). In a theta solvent, polymer–polymer and polymer–solvent interactions are counterbalanced and $\nu = 1/2$, leading to large conformational fluctuations and a maximally heterogeneous ensemble (72). A chain in a theta solvent behaves as a Flory random coil (73). In a good solvent, ν is ~3/5 and the chain is highly expanded due to favorable chain–solvent interactions. Under strongly denaturing conditions proteins show global dimensions consistent with a chain in a good solvent, while folded proteins can, on average, be described by a scaling relationship consistent with a chain in a poor solvent (13, 15, 16, 48). These scaling relationships describe the behavior of infinitely long homopolymers, although they are often used to characterize finite-sized heteropolymeric sequences (19, 27, 37, 38, 41, 45, 48, 53, 74–77). Consequently, for finite-sized heteropolymers we recast ν to be ν^{app}, which we define to be an apparent scaling exponent that best describes the apparent correlation length for a heteropolymer sequence of a specific, finite length that is at a certain distance from the theta temperature. We stress that the value of ν^{app} should not be conflated with well-established universal exponents, which are fixed points for infinitely long homopolymers.

We analyzed the scaling of interresidue distances as a function of sequence separation (Fig. 7*D*). This scaling behavior was compared with the scaling profile extracted from simulations of NTL9 in a bona fide theta solvent, i.e., Gaussian chain, in which all interresidue interactions are explicitly turned off and the intrinsic conformational preferences are modeled using a specific implementation of Flory's rotational isomeric approximation (27). Interestingly, despite the discernible presence of local structure and nonlocal interactions, the scaling profiles calculated for the unfolded-state ensemble in 1 M urea are consistent with that of a polymer in a theta solvent, with an apparent scaling exponent of $\nu^{app} = 0.48$ (Fig. 7*D*). This result implies that conformational fluctuations lead to a screening whereby the effects of intrachain attractions and repulsions counterbalance one another, on average, to yield an ensemble that has the statistical features of a finite-sized homopolymer in an apparent theta solvent. Fluctuations into and out of specific types of local secondary structures and the sampling of specific patterns of local and nonlocal contacts highlight the contributions of sequence-specific effects in the unfolded ensemble. The amplitudes of and correlations among conformational fluctuations within the ensemble will lead to a counterbalancing of intrachain repulsions by attractions and vice versa. As a result, the internal distance distributions and the overall dimensions end up in the same universality class as infinitely long homopolymers in theta solvents. In summary, the observed congruence between the Flory random coil and the unfolded-state ensemble in 1 M urea is a form of screening that is achievable in polymers whereby overcompensatory conformational fluctuations screen one another, leading to ensemble averages that are concordant with properties of canonical random coils.

We compared our results here with those from previous studies to show that the unfolded state of NTL9 under folding conditions undergoes a modest contraction as a function of decreasing urea concentration. We reach this conclusion by comparing the R_g values of 23.5 Å in 10 M urea and 21.3 Å in 8 M urea to the R_g value of 19.1 Å in 1 M urea (Fig. 8*A*) (48). The relative extent of contraction from high to low concentrations of urea is 19%. This level of contraction is consistent with values reported previously for ACTR (18%), R17 (23%), and ubiquitin (16%) (*SI Appendix*, Table S16) (9, 37, 38). The R_g values obtained for NTL9 in 10 M, 8 M, and 1 M urea are well described by simulations, from which additional parameters such as the scaling exponent and end-to-end distance can be directly calculated. For highly denatured states in 10 M and 8 M urea, the R_g value can also be accurately inferred from the end-to-end distance (Fig. 8*A*). However, for the unfolded state under folding conditions the R_g inferred from the end-to-end distance underestimates the true R_g. Similarly, if we estimate ν^{app} by extrapolating from the global dimensions assuming a fixed prefactor (*SI Appendix*, *SI Materials and Methods*) or by directly calculating from simulations, we obtain good agreement under denaturing conditions, but a range of values for the unfolded state under folding conditions (Fig. 8*B*). These results are consistent with a general model in which sequence-specific interactions introduce significant deviations from idealized homopolymer models under native conditions, whereby the end-to-end distance and R_g can become decoupled for finite-sized heteropolymers that are not approximated as bead-spring polymers. This raises a cautionary note about using generic conversions between R_g and end-to-end distance based on observations for homopolymers in theta or good solvents or relying on simulations that map proteins to equivalent homopolymers (53, 55, 56) or coarse-grained heteropolymers. In contrast, under strongly denaturing conditions, the strong and approximately sequence-independent solvent–chain interactions overpower sequence-specific intrachain interactions such that homopolymer models appear to be effective and appropriate models for describing polypeptides.

Discussion

Our work provides a high-resolution description of the unfolded state of NTL9 under refolding conditions. This is obtained through a combination of multiple, minimally invasive FRET probes in time-resolved FRET measurements, time-resolved SAXS, all-atom simulations, and polymer theory. Taken together, our results demonstrate that the unfolded state under refolding conditions is more compact than the fully denatured state, but still relatively expanded vis-à-vis the folded state. The unfolded state under refolding conditions is conformationally heterogeneous and is characterized by the formation of secondary and tertiary structures of the native and nonnative variety. Interresidue contacts are congruent with inferences from experiments, suggesting the importance of electrostatic and hydrophobic interactions in the unfolded state of NTL9 under highly denaturing and near native conditions (48, 78). The unfolded ensemble contains residual native helical structure in regions corresponding to the first and second α-helices. Experiments performed on peptide fragments corresponding to the first and second α-helices have shown that the second helix partially forms in isolation while the first one does not (79). Helical structure in the C-terminal helix thus requires

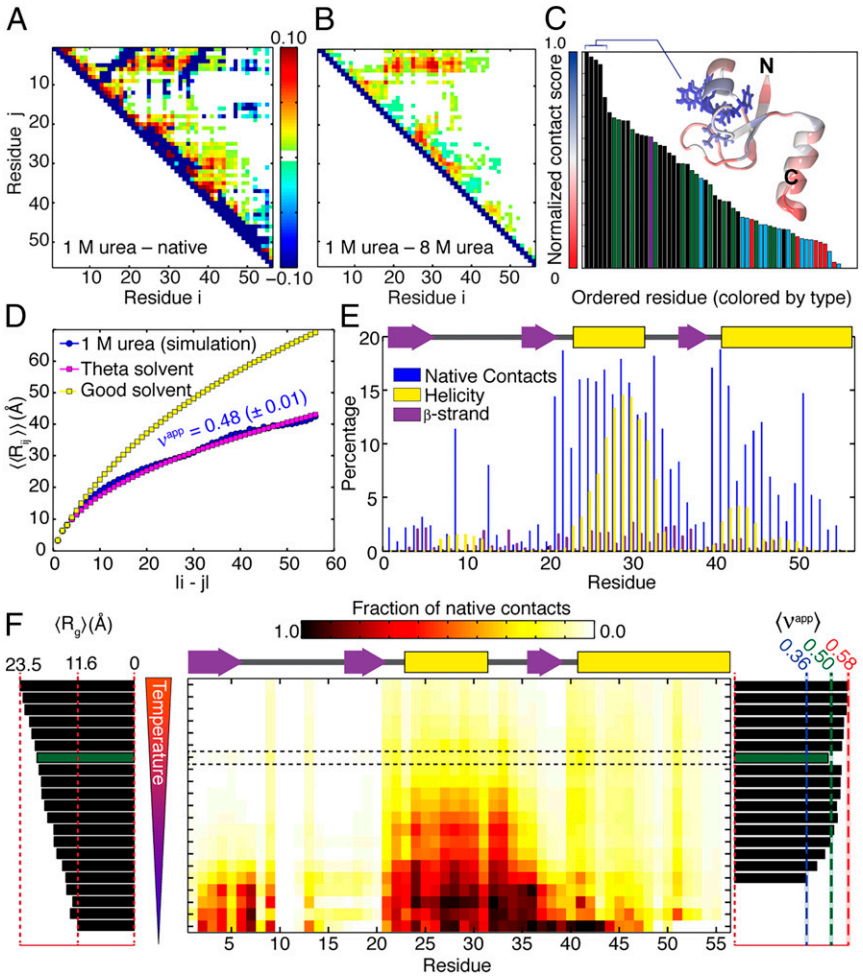

Fig. 7. Simulation results and analysis using polymer theory reveal the details of the unfolded state under refolding conditions. (*A*) Difference contact map generated by taking the 1-M urea ensemble contact map and subtracting the native-state contact map. Positive values correspond to nonnative interactions, while negative values reflect missing native contacts. (*B*) Difference contact map generated by taking the 1-M urea ensemble contact map and subtracting the 8-M urea contact map (48). Positive values correspond to contacts observed at 1 M urea that are not observed under strongly denaturing conditions. Specifically, we observe extensive contacts involving a cluster of N-terminal hydrophobic residues (V3, I4, F5, L6) and a cluster of central hydrophobic residues (F29, L30, F31). (*C*) The normalized per-residue contact score reflects the extent to which each residue is in direct contact with any other residue. Intramolecular interactions are driven by hydrophobic and aromatic residues, as shown by ordering all residues and coloring by their physicochemical properties (black, hydrophobic/aromatic; green, polar; purple, proline; blue, positive; red, negative). In particular, a cluster of hydrophobic residues (F31, F29, L30, Y25, L35) engages in extensive intramolecular interactions in the unfolded state and forms a hydrophobic cluster in the native structure. (*D*) Comparison of internal scaling profiles for NTL9 in a good solvent (yellow), in a theta solvent (pink), and from the 1-M urea ensemble (blue). In a globally averaged analysis, the 1-M urea ensemble shows scaling behavior consistent with a polymer in a theta solvent with an apparent scaling exponent of 0.48 ± 0.01, despite the presence of local structure and long-range interactions. (*E*) Presence of native contacts and secondary structure in the 1-M urea unfolded-state ensemble. The native-state secondary structure map is shown above for reference. (*F*) For denatured-state ensembles generated between 500 K and 280 K we quantify ensemble-averaged radius of gyration (*Left*), the residue-specific density of native contacts (*Center*), and apparent scaling exponent ν^{app} (*Right*). The reweighted ensemble is highlighted in the dashed box and with green bars.

only local contacts to form, while the first helix appears to be stabilized by tertiary interactions.

Udgaonkar and coworkers (80–83) have followed the evolution of multiple pairs of interresidue distances as a function of denaturant, using the quenching of Trp fluorescence by DTNB, and arrived at conclusions that support continuous contraction as a function of dilution from denaturant. Although the short R_0 of the F_{CN}-Trp pair may be considered a drawback for some applications, in the current study it offers the advantage of allowing accurate distances to be determined (Fig. 1) (53). The attachment of large fluorescent dyes to single-domain proteins has led to the oft-quoted concern that fluorophores must be inducing chain collapse (84). While the impact of dyes on the unfolded state will undoubtedly be a function of the physicochemical properties of the dyes and the specific sequences in question, our studies demonstrate that contraction of the unfolded state is observed even when using minimally perturbative probes—a finding that is concordant with observations made in single-molecule studies using larger dyes (19).

The distinctive nature of the unfolded state populated under native conditions is shown by the nonmonotonic relationship between sequence separation and spatial separation in Fig. 4. This behavior highlights the importance of probing multiple regions within the chain to obtain a more complete description of the ensemble (38). By combining multiple FRET pairs, each of which queries a relatively short distance, we have been able to construct a high-resolution and consistent description of the unfolded-state ensemble in 1 M urea. It has been proposed that hydrophobic clusters of Leu, Ile, Val, and Phe occur in the unfolded state because they are formed early in folding (85). Consistent with this hypothesis, we identified a specific cluster of residues that appear to form extensive native and nonnative contacts (Fig. 7*C*). Similarly, two clusters of hydrophobic residues separated in sequence by ∼25 residues engage in long-range contacts in the unfolded state. These residues also form a contiguous cluster in the folded core (residues 3–5 and residues 29–31) (Fig. 7*B*). Studies have reported that certain nonnative interactions can lower the free energy barrier for folding and increase the folding rate, that nonnative contacts can act to constrain the formation of native contacts, and that they may modulate folding rates without influencing the mechanism (86–91). Nonnative interactions involving hydrophobic residues may also reduce the likelihood of these residues forming intermolecular interactions that can lead the protein to aggregate.

Recent work has suggested that the unfolded state under folding conditions behaves as a polymer with an apparent scaling exponent of 0.54 (45). The unweighted simulated unfolded ensemble for NTL9 at 375 K (the ensemble that, before reweighting, most closely matched the FRET distances) has a scaling exponent of 0.55. However, upon reweighting to achieve congruence with the totality of the FRET data, the apparent scaling exponent drops to 0.48. This ensemble also reproduces the R_g value measured using SAXS. The apparent scaling exponent for the ensemble that matches all of the experimental data is actually in closer agreement to the value of 0.45 ± 0.03 from work by Hoffmann et al. (19). This is noteworthy because the scattering profiles calculated using the weighted and unweighted ensembles are essentially identical to one another

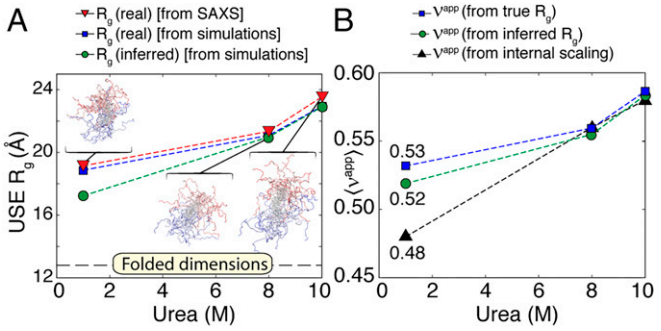

Fig. 8. The unfolded-state ensemble (USE) undergoes a continuous reduction in chain dimensions as a function of decreased denaturant concentration. (*A*) SAXS shows a continuous reduction in chain dimensions (red triangles), a trend that is matched by simulated unfolded-state ensembles. The ensemble-averaged R_g can be estimated from the end-to-end distance, assuming $R_g = \frac{R_e}{\sqrt{6}}$. This approximation works reasonably well under strongly denaturing conditions but underestimates the R_g under native conditions. (*B*) ν^{app} can be estimated from global dimensions or calculated directly from the unfolded-state ensemble. As with R_g, these approximations for ν^{app} work well under strongly denaturing conditions, but under native conditions a range of ν^{app} values are obtained.

(*SI Appendix*, Fig. S5*G*). Therefore, despite a minor change between the unweighted and reweighted ensembles, we observe a shift of ∼0.07 in the apparent scaling exponent. This observation highlights the fact that relying exclusively on scattering profiles can yield incomplete insights because the overall profile is insensitive to the effects of fluctuations and low-likelihood contacts. These can be gleaned only by directly probing contacts or preferred distances across multiple length scales without the confounding effects of averaging across all of the distances that happen in SAXS. In the context of our studies, values of 0.48 and 0.55 for the apparent scaling exponent are almost indistinguishable from one another as evaluated by different metrics for probing interresidue contacts and distances (Fig. 8). Moreover, differences in the methods used for estimating ν^{app} from simulations can introduce differences of ±0.05, while assumptions made for estimating this value using experimental data can have similar consequences. Our results here and elsewhere paint a picture in which polypeptides with ν^{app} between 0.45 and 0.55 are consistent with expanded, heterogeneous ensembles with ensemble-averaged properties being congruent with polymers in theta solvents while supporting the presence of extensive, sequence-specific local and long-range interactions. We propose that it is erroneous to assert very precise universal values for ν^{app} because the inferred value for ν^{app} is influenced by the resolution of the experimental techniques used and by underlying assumptions regarding polymer behavior that one makes to analyze the acquired data. In a similar vein, while we (and others) have found a modest contraction of the unfolded-state ensemble from high to low denaturant of around 20%, we do not necessarily expect this value to hold true for every protein, since sequence-specific effects will determine the consequences of the interplay between intrachain and chain–solvent interactions, an expectation borne out in numerous experimental examples (19, 27, 37, 41, 53, 92–94).

Conclusions

The combination of multiple time-resolved FRET, time-resolved SAXS, and computational results shows that the unfolded state under folding conditions is conformationally heterogeneous; certain regions of the protein undergo significant collapse while others remain relatively expanded. The overall ensemble-averaged properties such as the scaling of internal distances, the dimensions of the unfolded state under folding conditions, and the inferred apparent scaling exponent are concordant with properties of polymers in theta solvent. Similar results have been reported for several other proteins that were studied using single-molecule FRET and/or SAXS (37, 38, 95). However, the observation that the global dimensions of the unfolded ensemble for NTL9 in 1 M urea resemble those of chains in theta solvents does not mean that unfolded ensembles are ideal chains devoid of long-range interactions. The Flory-style models provide a way to describe the endpoint of conformational averaging and they say nothing about how the averaging actually comes about. Further, the results presented here highlight the difficulty in using inferred apparent scaling exponents as the sole device for defining the properties of unfolded states. Our results suggest that the apparent scaling exponent of $\nu^{app} = 0.48$ is the result of counterbalancing the effects of intrachain repulsions. This is evident in the observation of segment-specific expansion coexisting with segment-specific contraction. The former should result from intrachain repulsions whereas the latter should result from intrachain attractions. It seems plausible that unfolded states characterized by amplitudes of and correlations among conformational fluctuations that are reminiscent of polymers in theta states may facilitate the search for the folded state more effectively than an expanded self-avoiding random walk or compact globular ensemble that hinders conformational diffusion (16, 19, 72, 96, 97). Our ability to uncover atomistic descriptions of unfolded ensembles under folding conditions opens the door to designing features into unfolded states without impacting the properties of folded states to interrogate the effects of such designs on folding, function, and cellular phenotypes.

Materials and Methods

Preparation and Characterization of *p*-cyanoPhe-Labeled NTL9. Seven distinct variants of NTL9 containing F_{CN} and Trp were designed and generated. Matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry was used to confirm the molecular weight of each species (*SI Appendix*, Fig. S8). Variants containing the donor only and the donor in the presence of acceptor were prepared recombinantly using 21st pair technology developed by the Schultz and Mehl laboratories. All of the mutants are folded as judged by CD, are minimally destabilized compared with the wild-type protein, and display sigmoidal unfolding transitions (*SI Appendix*, Fig. S1 and Tables S1–S4). The stability measurements demonstrate that F_{CN} is a minimally perturbative substitution in NTL9, not only for Tyr and Phe, but also for Lys and Gln. Only one of the variants, N43F_{CN}, is destabilized by >1 kcal/mol relative to wild type; however, the CD spectrum of this mutant is very similar to that of wild-type NTL9, indicating that overall native secondary structure is not perturbed. Observed relaxation rates in 1 M urea were determined for the variants from continuous-flow experiments and do not deviate significantly from that of the wild-type protein (*SI Appendix*, Fig. S3 and Table S5).

Analysis of Time-Resolved FRET Data. The folded-state fluorescence lifetime data were fitted using a Gaussian distribution, where the mean distances range from 8 Å for the 33F_{CN}Trp25 pair to 21 Å for the 42F_{CN}Trp25 and 25F_{CN}Trp33 pairs. The high-denaturant unfolded state was fitted separately using a Gaussian distribution and a WLC model, taking into account diffusion during the excited-state lifetime (*SI Appendix*, Tables S9 and S10). Both models have been used to fit FRET data, but our results here show no dependence on which of the models is used, providing confidence that the underlying model is not biasing the derived distances (51, 52, 98).

SVD analysis of the continuous-flow data in 1 M urea gives a maximum of two components along both the folding and the time-correlated single-photon counting (TCSPC) axis. The amplitude of the major and minor SVD components along the folding time axis fitted well to a single exponential model, consistent with two-state folding (*SI Appendix*, Fig. S3). This allowed us to globally analyze the continuous-flow data using a two-population model, folded and unfolded. The observation that the data are well fitted using two SVD components does not mean that the unfolded-state properties are independent of urea, and clear differences between the 10-M and 1-M unfolded states are observed. As before, the unfolded states were fitted using a Gaussian distribution and the WLC model with diffusion. The WLC model contains only the amplitude and persistence length as fitting parameters and this allows a diffusion parameter to be introduced without overparameterizing the fits. Although the absolute value of the distances is slightly different between the Gaussian distribution and WLC models, fits to both models with and without diffusion indicate significant compaction of the protein upon dilution out of high denaturant. The greatest extent of collapse is seen for the two variants that probe the largest sequence

separation, 2–25 and 2–33 (Fig. 3). The mean distance in the unfolded state derived from WLC analysis of the F_{CN}2–Trp25 pair is 36.5 Å in 10 M urea and 21.4 Å in 1 M urea, while the mean distances observed for the F_{CN}2–Trp33 pair are 40.1 Å in 10 M urea and 24.1 Å in 1 M urea. For additional information on the details associated with the fitting and interpretation of FRET data, please see *SI Appendix*.

Recording and Analysis of Time-Resolved SAXS Data. Four refolding experiments were performed and averaged in the analysis. The earliest time point included in the analysis was 200 μs. Scattering curves as a function of refolding time were fitted globally to a two-state kinetic model and extrapolated to time = 0 and time = ∞ to determine the scattering curve for the unfolded and native states, respectively. The population of folded molecules in 8 M urea was accounted for in the analysis. The pair distribution function, P(r), for the folded and unfolded states in 1 M urea was also determined (*SI Appendix*, Fig. S5).

Generation, Reweighting, and Analysis of Denatured-State Ensembles. Starting from the NTL9 crystal structure we generated 22 unconstrained ensembles for a range of temperatures (240–500 K) and then filtered out conformations where more than 50% native contacts were observed to ensure we had true "denatured-state" ensembles (*SI Appendix*, Table S14). To generate each ensemble, 10 independent simulations were run for 8×10^7 Monte Carlo steps, with the first 2×10^7 steps discarded as equilibration. To ensure the reproducibility and convergence of simulations, we compared two sets of 40 simulations, with each set started either from the folded state or from a randomly generated unfolded state (80 simulations total). The ensemble averages computed across the different sets were similar to one another, thus highlighting the high quality and convergence of Monte Carlo importance sampling (*SI Appendix*, Fig. S9).

Polymer Scaling Analysis in Finite Chains. The apparent scaling exponent (ν^{app}) was estimated by fitting the results of simulations using the following expression:

$$\ln\left(\sqrt{\langle r_{ij}\rangle^2}\right) = \nu^{app} \ln(|i-j|) + A_0. \quad [2]$$

All sequence separations, $|i - j|$, that are greater than 15 and up to n_{res} – 5 were used. Here, n_{res} is the number of residues in NTL9. The lower bound of 15 corrects for deviation from scaling behavior at short distances, while the upper bound of n_{res} – 5 corrects for the effects "dangling ends" that introduce deviations from the infinitely long chain limit (the limit under which scaling theory was developed). Points were distributed such that they are evenly spaced in log–log space, avoiding common pitfalls associated with fitting linear models to log–log plots.

ACKNOWLEDGMENTS. This work was supported by Wellcome Trust Grant 107927/Z/15/Z (to D.P.R.), the US National Science Foundation (NSF) Grants IDBR 1353942 and MCB 0721312 (to O.B.) and MCB-1614766 (to R.V.P.), and the Human Frontiers Science Program Grant RGP0034/2017 (to R.V.P.). I.P. was supported in part by NSF Grant MCB-1330259 (to D.P.R.). O.B., R.V.P., and D.P.R. are members of the protein-folding consortium that is supported by the NSF through Grant MCB 1051344. We thank S. Chakravarthy and S. V. Kathuria for assistance with SAXS measurements. We are grateful to current and former members of the O.B., R.V.P., and D.P.R. laboratories as well as Prof. C. R. Matthews for helpful discussions. This research used resources of the Advanced Photon Source, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract DE-AC02-06CH11357. This project was supported by Grant 9 P41 GM103622 from the National Institute of General Medical Sciences of the National Institutes of Health.

1. C. Levinthal, Are there pathways for protein folding? *J. Chim. Phys.* **65**, 44–45 (1968).
2. M. Karplus, D. L. Weaver, Protein-folding dynamics. *Nature* **260**, 404–406 (1976).
3. K. A. Dill, Dominant forces in protein folding. *Biochemistry* **29**, 7133–7155 (1990).
4. K. A. Dill, J. L. MacCallum, The protein-folding problem, 50 years on. *Science* **338**, 1042–1046 (2012).
5. H. S. Chung, S. Piana-Agostinetti, D. E. Shaw, W. A. Eaton, Structural origin of slow diffusion in protein folding. *Science* **349**, 1504–1510 (2015).
6. K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
7. V. A. Voelz, G. R. Bowman, K. Beauchamp, V. S. Pande, Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J. Am. Chem. Soc.* **132**, 1526–1528 (2010).
8. C. Magg, F. X. Schmid, Rapid collapse precedes the fast two-state folding of the cold shock protein. *J. Mol. Biol.* **335**, 1309–1323 (2004).
9. J. Jacob, B. Krantz, R. S. Dothager, P. Thiyagarajan, T. R. Sosnick, Early collapse is not an obligate step in protein folding. *J. Mol. Biol.* **338**, 369–382 (2004).
10. T. R. Sosnick, D. Barrick, The folding of single domain proteins–Have we reached a consensus? *Curr. Opin. Struct. Biol.* **21**, 12–24 (2011).
11. T. Y. Yoo *et al.*, Small-angle X-ray scattering and single-molecule FRET spectroscopy produce highly divergent views of the low-denaturant unfolded state. *J. Mol. Biol.* **418**, 226–236 (2012).
12. G. Haran, How, when and why proteins collapse: The relation to folding. *Curr. Opin. Struct. Biol.* **22**, 14–20 (2012).
13. J. E. Kohn *et al.*, Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 12491–12496 (2004).
14. H. T. Tran, A. Mao, R. V. Pappu, Role of backbone-solvent interactions in determining conformational equilibria of intrinsically disordered proteins. *J. Am. Chem. Soc.* **130**, 7380–7392 (2008).
15. R. I. Dima, D. Thirumalai, Asymmetry in the shapes of folded and denatured states of proteins. *J. Phys. Chem. B* **108**, 6564–6570 (2004).
16. A. S. Holehouse, R. V. Pappu, Collapse transitions of proteins and the interplay among backbone, sidechain, and solvent interactions. *Annu. Rev. Biophys.* **47**, 19–39 (2018).
17. E. A. Lipman, B. Schuler, O. Bakajin, W. A. Eaton, Single-molecule measurement of protein folding kinetics. *Science* **301**, 1233–1235 (2003).
18. M. Brucale, B. Schuler, B. Samorì, Single-molecule studies of intrinsically disordered proteins. *Chem. Rev.* **114**, 3281–3317 (2014).
19. H. Hofmann *et al.*, Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16155–16160 (2012).
20. B. Schuler, E. A. Lipman, W. A. Eaton, Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature* **419**, 743–747 (2002).
21. A. A. Deniz *et al.*, Single-molecule protein folding: Diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5179–5184 (2000).
22. E. Sherman, G. Haran, Coil-globule transition in the denatured state of a small protein. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 11539–11543 (2006).
23. T. R. Jahn, S. E. Radford, Folding versus aggregation: Polypeptide conformations on competing pathways. *Arch. Biochem. Biophys.* **469**, 100–117 (2008).
24. D. Eisenberg, M. Jucker, The amyloid state of proteins in human diseases. *Cell* **148**, 1188–1203 (2012).
25. V. N. Uversky, Natively unfolded proteins: A point where biology waits for physics. *Protein Sci.* **11**, 739–756 (2002).
26. H. J. Dyson, P. E. Wright, Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).
27. A. S. Holehouse, K. Garai, N. Lyle, A. Vitalis, R. V. Pappu, Quantitative assessments of the distinct contributions of polypeptide backbone amides versus side chain groups to chain expansion via chemical denaturation. *J. Am. Chem. Soc.* **137**, 2984–2995 (2015).
28. R. K. Das, K. M. Ruff, R. V. Pappu, Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **32**, 102–112 (2015).
29. H. S. Chan, K. A. Dill, Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Biophys. Chem.* **20**, 447–490 (1991).
30. H. Roder, K. Maki, H. Cheng, Early events in protein folding explored by rapid mixing methods. *Chem. Rev.* **106**, 1836–1861 (2006).
31. O. Bilsel, C. R. Matthews, Molecular dimensions and their distributions in early folding intermediates. *Curr. Opin. Struct. Biol.* **16**, 86–93 (2006).
32. Y. Wu, E. Kondrashkina, C. Kayatekin, C. R. Matthews, O. Bilsel, Microsecond acquisition of heterogeneous structure in the folding of a TIM barrel protein. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 13367–13372 (2008).
33. T. Kimura *et al.*, Specific collapse followed by slow hydrogen-bond formation of beta-sheet in the folding of single-chain monellin. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 2748–2753 (2005).
34. S. V. Kathuria *et al.*, Microsecond barrier-limited chain collapse observed by time-resolved FRET and SAXS. *J. Mol. Biol.* **426**, 1980–1994 (2014).
35. M. Sadqi, L. J. Lapidus, V. Muñoz, How fast is protein hydrophobic collapse? *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12117–12122 (2003).
36. M. Arai, M. Iwakura, C. R. Matthews, O. Bilsel, Microsecond subdomain folding in dihydrofolate reductase. *J. Mol. Biol.* **410**, 329–342 (2011).
37. A. Borgia *et al.*, Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. *J. Am. Chem. Soc.* **138**, 11714–11726 (2016).
38. M. Aznauryan *et al.*, Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E5389–E5398 (2016).
39. T. Tezuka-Kawakami, C. Gell, D. J. Brockwell, S. E. Radford, D. A. Smith, Urea-induced unfolding of the immunity protein Im9 monitored by spFRET. *Biophys. J.* **91**, L42–L44 (2006).
40. F. Huang, L. Ying, A. R. Fersht, Direct observation of barrier-limited folding of BBL by single-molecule fluorescence resonance energy transfer. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 16239–16244 (2009).
41. W. Zheng *et al.*, Probing the action of chemical denaturant on an intrinsically disordered protein by simulation and experiment. *J. Am. Chem. Soc.* **138**, 11702–11713 (2016).
42. G. Raos, G. Allegra, Macromolecular clusters in poor-solvent polymer solutions. *J. Chem. Phys.* **107**, 6479–6490 (1997).
43. C. J. Camacho, D. Thirumalai, Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 6369–6372 (1993).

44. G. L. Hura *et al*., Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat. Methods* **6**, 606–612 (2009).
45. J. A. Riback *et al*., Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **358**, 238–241 (2017).
46. G. Tria, H. D. T. Mertens, M. Kachala, D. I. Svergun, Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ* **2**, 207–217 (2015).
47. A. G. Kikhney, D. I. Svergun, A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.* **589**, 2570–2577 (2015).
48. W. Meng, N. Lyle, B. Luan, D. P. Raleigh, R. V. Pappu, Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 2123–2128 (2013).
49. B. Luan, N. Lyle, R. V. Pappu, D. P. Raleigh, Denatured state ensembles with the same radii of gyration can form significantly different long-range contacts. *Biochemistry* **53**, 39–47 (2014).
50. B. Schuler, A. Soranno, H. Hofmann, D. Nettels, Single-molecule FRET spectroscopy and the polymer physics of unfolded and intrinsically disordered proteins. *Annu. Rev. Biophys.* **45**, 207–231 (2016).
51. E. P. O'Brien, G. Morrison, B. R. Brooks, D. Thirumalai, How accurate are polymer models in the analysis of Förster resonance energy transfer experiments on proteins? *J. Chem. Phys.* **130**, 124903 (2009).
52. J. Song, G.-N. Gomes, C. C. Gradinaru, H. S. Chan, An adequate account of excluded volume is necessary to infer compactness and asphericity of disordered proteins by Förster resonance energy transfer. *J. Phys. Chem. B* **119**, 15191–15202 (2015).
53. G. Fuertes *et al*., Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E6342–E6351 (2017).
54. J. B. Warner, 4th *et al*., Monomeric huntingtin exon 1 has similar overall structural features for wild-type and pathological polyglutamine lengths. *J. Am. Chem. Soc.* **139**, 14456–14469 (2017).
55. K. M. Ruff, A. S. Holehouse, SAXS versus FRET: A matter of heterogeneity? *Biophys. J.* **113**, 971–973 (2017).
56. J. Song, G. N. Gomes, T. Shi, C. C. Gradinaru, H. S. Chan, Conformational heterogeneity and FRET data interpretation for dimensions of unfolded proteins. *Biophys. J.* **113**, 1012–1024 (2017).
57. G. Fuertes *et al*., Comment on "Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water." *Science* **361**, eaau8230 (2018).
58. R. B. Best *et al*., Comment on "Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water." *Science* **361**, eaar7101 (2018).
59. J. A. Riback *et al*., Response to comment on "Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water." *Science* **361**, eaar7949 (2018).
60. J. A. Riback, *et al*., Commonly used FRET fluorophores promote collapse of an otherwise disordered protein. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 8889–8894 (2019).
61. D. W. Hoffman *et al*., Crystal structure of prokaryotic ribosomal protein L9: A bi-lobed RNA-binding protein. *EMBO J.* **13**, 205–212 (1994).
62. B. Kuhlman, D. L. Luisi, P. A. Evans, D. P. Raleigh, Global analysis of the effects of temperature and denaturant on the folding and unfolding kinetics of the N-terminal domain of the protein L9. *J. Mol. Biol.* **284**, 1661–1670 (1998).
63. B. Anil, S. Sato, J. H. Cho, D. P. Raleigh, Fine structure analysis of a protein folding transition state; distinguishing between hydrophobic stabilization and specific packing. *J. Mol. Biol.* **354**, 693–705 (2005).
64. M. J. Tucker, R. Oyola, F. Gai, Conformational distribution of a 14-residue peptide in solution: A fluorescence resonance energy transfer study. *J. Phys. Chem. B* **109**, 4788–4795 (2005).
65. L. Wang, J. Xie, P. G. Schultz, Expanding the genetic code. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 225–249 (2006).
66. S. J. Miyake-Stoner *et al*., Probing protein folding using site-specifically encoded unnatural amino acids as FRET donors with tryptophan. *Biochemistry* **48**, 5953–5962 (2009).
67. O. Glatter, O. Kratky, *Small Angle X-Ray Scattering* (Academic Press, London, 1982).
68. H. Boze *et al*., Proline-rich salivary proteins have extended conformations. *Biophys. J.* **99**, 656–665 (2010).
69. R. K. Das, Y. Huang, A. H. Phillips, R. W. Kriwacki, R. V. Pappu, Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 5616–5621 (2016).
70. H. T. A. Leung *et al*., A rigorous and efficient method to reweight very large conformational ensembles using average experimental data and to determine their relative information content. *J. Chem. Theory Comput.* **12**, 383–394 (2016).
71. P. G. De Gennes, Collapse of a polymer chain in poor solvents. *J. Physique Lett.* **36**, 55–57 (1975).
72. N. Lyle, R. K. Das, R. V. Pappu, A quantitative measure for protein conformational heterogeneity. *J. Chem. Phys.* **139**, 121907 (2013).
73. A. H. Mao, N. Lyle, R. V. Pappu, Describing sequence-ensemble relationships for intrinsically disordered proteins. *Biochem. J.* **449**, 307–318 (2013).
74. A. Soranno *et al*., Single-molecule spectroscopy reveals polymer effects of disordered proteins in crowded environments. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 4874–4879 (2014).
75. A. Soranno *et al*., Quantifying internal friction in unfolded and intrinsically disordered proteins with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17800–17806 (2012).
76. S. Müller-Späth *et al*., From the cover: Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 14609–14614 (2010).
77. P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, NY, 1953).
78. J.-H. Cho *et al*., Energetically significant networks of coupled interactions within an unfolded protein. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 12079–12084 (2014).
79. B. Kuhlman, D. L. Luisi, P. Young, D. P. Raleigh, pKa values and the pH dependent stability of the N-terminal domain of L9 as probes of electrostatic interactions in the denatured state. Differentiation between local and nonlocal interactions. *Biochemistry* **38**, 4896–4903 (1999).
80. J. B. Udgaonkar, Polypeptide chain collapse and protein folding. *Arch. Biochem. Biophys.* **531**, 24–33 (2013).
81. A. Dasgupta, J. B. Udgaonkar, Evidence for initial non-specific polypeptide chain collapse during the refolding of the SH3 domain of PI3 kinase. *J. Mol. Biol.* **403**, 430–445 (2010).
82. K. K. Sinha, J. B. Udgaonkar, Dissecting the non-specific and specific components of the initial folding reaction of barstar by multi-site FRET measurements. *J. Mol. Biol.* **370**, 385–405 (2007).
83. S. Bhatia, G. Krishnamoorthy, J. B. Udgaonkar, Site-specific time-resolved FRET reveals local variations in the unfolding mechanism in an apparently two-state protein unfolding transition. *Phys. Chem. Chem. Phys.* **20**, 3216–3232 (2018).
84. J. A. Riback *et al*., Saxs confirms that FRET dyes promote collapse of an otherwise fully disordered protein. *Biophys. J.* **114**, 368A (2018).
85. S. V. Kathuria, Y. H. Chan, R. P. Nobrega, A. Özen, C. R. Matthews, Clusters of isoleucine, leucine, and valine side chains define cores of stability in high-energy states of globular proteins: Sequence determinants of structure and stability. *Protein Sci.* **25**, 662–675 (2016).
86. P. F. Faísca, A. Nunes, R. D. Travasso, E. I. Shakhnovich, Non-native interactions play an effective role in protein folding dynamics. *Protein Sci.* **19**, 2196–2209 (2010).
87. R. B. Best, G. Hummer, Coordinate-dependent diffusion in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 1088–1093 (2010).
88. S. S. Plotkin, Speeding protein folding beyond the G(o) model: How a little frustration sometimes helps. *Proteins* **45**, 337–345 (2001).
89. C. Clementi, S. S. Plotkin, The effects of nonnative interactions on protein folding rates: Theory and simulation. *Protein Sci.* **13**, 1750–1766 (2004).
90. C. J. Camacho, D. Thirumalai, Modeling the role of disulfide bonds in protein folding: Entropic barriers and pathways. *Proteins* **22**, 27–40 (1995).
91. B. C. Gin, J. P. Garrahan, P. L. Geissler, The limited role of nonnative contacts in the folding pathways of a lattice protein. *J. Mol. Biol.* **392**, 1303–1314 (2009).
92. H. S. Samanta *et al*., Protein collapse is encoded in the folded state architecture. *Soft Matter* **13**, 3622–3638 (2017).
93. E. W. Martin *et al*., Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation. *J. Am. Chem. Soc.* **138**, 15323–15335 (2016).
94. G. Reddy, D. Thirumalai, Collapse precedes folding in denaturant-dependent assembly of ubiquitin. *J. Phys. Chem. B* **121**, 995–1009 (2017).
95. Y. Wang, J. Trewhella, D. P. Goldenberg, Small-angle X-ray scattering of reduced ribonuclease A: Effects of solution conditions and comparisons with a computational model of unfolded proteins. *J. Mol. Biol.* **377**, 1576–1592 (2008).
96. D. K. Klimov, D. Thirumalai, Factors governing the foldability of proteins. *Proteins* **26**, 411–441 (1996).
97. D. K. Klimov, D. Thirumalai, Criterion that determines the foldability of proteins. *Phys. Rev. Lett.* **76**, 4070–4073 (1996).
98. B. Y. Ha, D. Thirumalai, Semiflexible chains under tension. *J. Chem. Phys.* **106**, 4243–4247 (1997).