

Site-specific Dimensions Across a Highly Denatured Protein; A Single Molecule Study

Evan R. McCarney¹, James H. Werner², Summer L. Bernstein¹
Ingo Ruczinski³, Dmitrii E. Makarov⁴, Peter M. Goodwin² and
Kevin W. Plaxco^{1*}

¹Department of Chemistry and Biochemistry, University of California, Santa Barbara CA 93106, USA

²Bioscience Division, Los Alamos National Laboratory Mail Stop M888, Los Alamos NM 87545, USA

³Department of Biostatistics Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA

⁴Department of Chemistry and Biochemistry and Institute for Theoretical Chemistry University of Texas, Austin TX 78712, USA

Do highly denatured proteins adopt random coil configurations? Here, we address this question by measuring residue-to-residue separations across the denatured FynSH3 domain. Using single-molecule Förster resonance energy transfer techniques, we have collected transfer efficiency probability distributions for dye-labeled, denatured protein. Applying maximum likelihood analysis to the interpretation of these distributions, we have determined the through-space distance between five residue pairs in the protein's guanidine hydrochloride-unfolded and trifluoroethanol-unfolded states. We find that, while the dimensions of the guanidine hydrochloride-unfolded molecule generally coincide with the dimensions predicted for a random coil ensemble, potentially statistically significant deviations from random coil behavior are also evident. These small, site-specific deviations may provide a means of reconciling earlier, scattering-based evidence for the random coil nature of the unfolded state with more site-specific spectroscopic evidence suggesting residual structure. We have also studied the unfolded ensemble populated in 50% trifluoroethanol, a denaturant that induces a highly helical unfolded state. We find that the size and shape of the unfolded ensemble under these conditions is effectively indistinguishable from that populated in guanidinium hydrochloride solutions, suggesting that the gross structure of the denatured state is, perhaps surprisingly, independent of the chemistry of the cosolvent.

© 2005 Published by Elsevier Ltd.

Keywords: protein folding; lattice polymer simulations; denatured state; random coil; single molecule

*Corresponding author

Introduction

Do denatured proteins behave as random coils? Spectroscopic studies suggest that proteins retain significant residual structure under even highly denaturing conditions. For example, numerous NMR studies provide evidence for non-random sequence-local and long-range structure even in the presence of high levels of the chemical denaturants urea and guanidine (GuHCl).^{1–7} Small-angle X-ray scattering (SAXS), in contrast, indicates that chemi-

cally denatured proteins adopt a random coil configuration; SAXS profiles are consistent with a Gaussian distribution of conformations,⁸ and the dimensions of only two of more than 24 denatured proteins deviate significantly from expected random coil scaling.⁹

The apparent discrepancy between the results of the SAXS and the NMR studies may be reconciled *via* the observation that the former reports on the average behavior of the entire polymer chain. In contrast, NMR reports on site-specific interactions,^{10–15} which, when averaged over the entire chain, might produce apparently random coil behavior.¹⁶ Here, we investigate this question in more detail. We do so by testing for site-specific, long-range deviations from random coil dimensions using single-pair Förster resonance energy transfer (FRET), a technique that allows us to

Abbreviations used: FRET, Förster resonance energy transfer; smFRET, single-molecule FRET; GuHCl, guanidine hydrochloride; TFE, trifluoroethanol; SAXS, small-angle X-ray scattering.

E-mail address of the corresponding author: kwp@chem.ucsb.edu

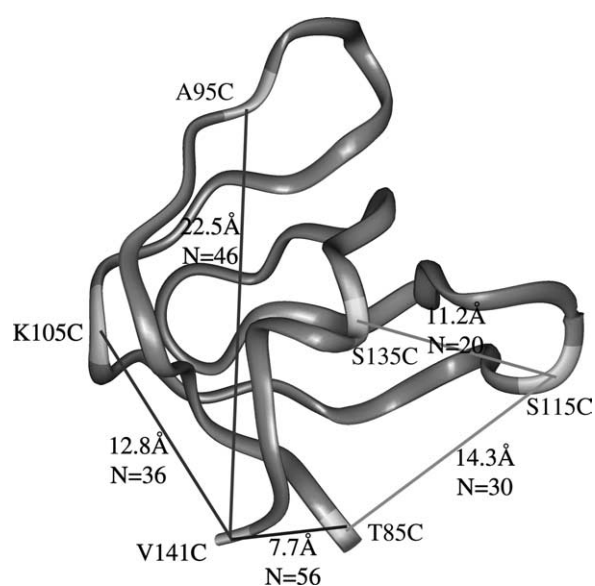


Figure 1. Five FynSH3 constructs were engineered, with cysteine pairs separated by 20, 30, 36, 46 or 56 amino acid residues. The cysteine sites are indicated and are connected by lines that denote pairs present in individual constructs. The through-space, native-state distance and sequence separation of each construct are indicated.

determine distances between specific residue pairs separated by 20–100 Å. Using five site-specifically labeled FynSH3 constructs (Figure 1), we have measured mean through-space distances between pairs of amino acid residues in the denatured polypeptide separated by 20–56 residues. When the data are compared to a derived random coil model, a high-resolution picture of the unfolded state is produced for this well-characterized protein.

Results

As a test system for our studies of the unfolded state, we have employed the FynSH3 domain, a single-domain, predominantly β -sheet protein that has been the subject of exhaustive kinetic and thermodynamic studies.¹⁷ We have studied the denatured states of this protein induced by both 4 M guanidine hydrochloride (GuHCl) and 50% (v/v) trifluoroethanol (TFE). Under these conditions, the characteristic positive ellipticity observed at 220 nm for native SH3 domains is entirely lost, suggesting that the protein is fully unfolded (Figure 2). Consistent with this, the CD spectra of GuHCl-denatured FynSH3 shows only very limited ellipticity, suggesting that the chain adopts random coil ϕ/ψ preferences under these conditions. The CD spectrum of the TFE-denatured protein, in contrast, exhibits a large, negative dip at 222 nm, which is indicative of significant, non-native helical content.

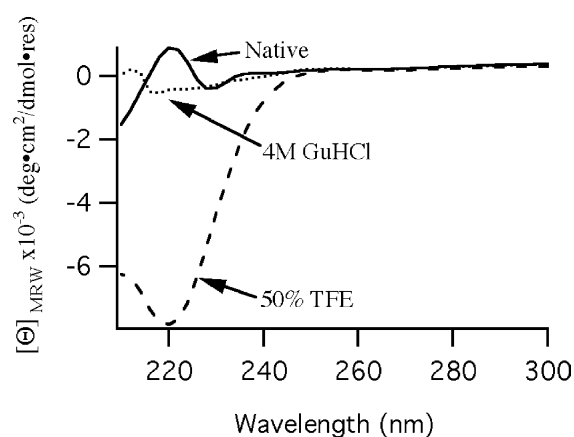


Figure 2. We have denatured FynSH3 using both 4 M GuHCl and 50% TFE. Under these conditions, the characteristic positive ellipticity observed at 220 nm for native SH3 domains is entirely lost, suggesting that the proteins are fully denatured. Consistent with this, the CD spectra of GuHCl-denatured FynSH3 shows only very limited ellipticity, suggesting that the chain adopts random coil ϕ/ψ preferences under these conditions. The CD spectrum of the TFE-denatured protein, in contrast, exhibits the large, negative dip at 222 nm that is indicative of significant, non-native helical content.

The determination of mean residue-to-residue distances *via* smFRET

FRET is the classic “spectroscopic ruler” of molecular biophysics. At the single-molecule level, however, FRET has generally been employed only rather qualitatively (i.e. true distances are calculated rarely or only approximately).^{18–23} Recent advances, however, have demonstrated the quantitative potential of single-molecule FRET (smFRET) techniques.^{24,25} Here, we address the many considerations necessary to quantitatively map FRET measurements into accurate mean distances. (We note also that, ultimately, such single-molecule studies will enable us to detect conformational heterogeneity in the unfolded state. To date, however, this has not proven feasible, due to the rapidity with which the unfolded ensemble averages relative to the millisecond observation times required for current generation single-molecule studies.²⁴)

The distance separating two fluorophores, R , is related to E , the “true” transfer efficiency, *via* the relationship $E = [1 + (R/R_0)^6]^{-1}$, where R_0 is a characteristic distance for a given donor–acceptor pair, termed the Forster radius (see below). Provided the estimates for R_0 and E are reliable, one can use this relationship to calculate inter-dye distances. Experimental studies, however, measure apparent energy transfer efficiencies, E_{app} , which can be calculated directly from the output of the donor (I_d) and acceptor (I_a) channels:

$$E_{app} = \frac{I_a}{I_a + I_d} \quad (1)$$

This differs from E , which is given by:

$$E = \left[1 + \gamma \frac{I_d - b_d}{I_a - b_a} \right]^{-1} \quad (2)$$

where the constant γ corrects for the donor and acceptor quantum yields, and the different collection efficiencies of the two channels,¹⁸ and b_d and b_a are the background of each channel. Here, we assume $\gamma=1$. This assumption is reasonable, because γ is known to be only slightly less than unity,¹⁸ and the distance between the donor and acceptor dyes, which is the value we wish to determine, is a function of $\gamma^{1/6}$, which approaches unity still more closely. E_{app} values calculated using

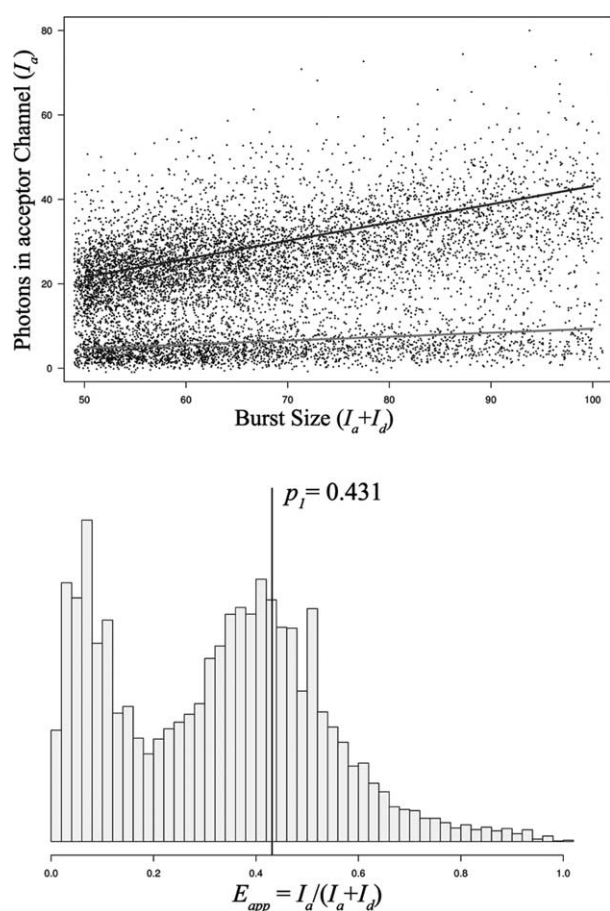


Figure 3. The energy transfer probability for the construct labeled at residues 85 and 141 as measured in 4 M GuHCl (for remaining data, see Supplementary Data). Top: for each burst, the number of photoelectrons detected in the acceptor channel, x_i , is plotted against total burst size, n_i , and fit with maximum likelihood estimation to give the fit lines $p_i \cdot n_i$ (p_0 , gray line and p_1 black line). This method allows us to distinguish between properly labeled molecules and mislabeled or photo-bleached molecules that would otherwise reduce the apparent energy transfer and thus systematically increase our estimate of donor–acceptor distance. Bottom: the data can be plotted as a traditional histogram of E_{app} probabilities. Two separable populations can be seen, one stemming from properly heterolabeled molecules and the other from incorrectly homolabeled proteins.

equation (1) agree with values calculated from lifetime measurements (data not shown), further supporting this assumption.

E_{app} is traditionally calculated for the burst of photons that is produced when a single molecule transits through the excitation volume. In previous smFRET studies, the E_{app} calculated for individual bursts have been plotted as a histogram and fitted to mixtures of normal, lognormal or Gaussian distributions (Figure 3, bottom).^{19,23–26} This approach, however, suffers from two potential shortcomings. First, E_{app} can take values only between zero and 1 (and a substantial number of points lie at the extremes), but it is fitted to distributions that extend beyond this range. This leads to potentially non-negligible departures from the assumptions about the distribution. Second, this approach weights all observations equally. Estimates stemming from larger bursts, however, are more precise and would thus, ideally, be weighted more heavily. Here, we account for these issues by employing a maximum likelihood estimation technique to determine the mean E_{app} , $\langle E_{app} \rangle$, for each construct.

We have measured FRET efficiency for thousands of single-molecule events for each of five FynSH3 constructs in 4 M GuHCl. Using maximum likelihood methods (Figure 3), we have determined $\langle E_{app} \rangle$ for each construct (Figure 4). Using our knowledge of the Forster radius (R_0) (see below), we can convert these values into effective distances, R_{eff} , via the equation:

$$\langle E_{app} \rangle = [1 + (R_{eff}/R_0)^6]^{-1} \quad (3)$$

The observed distances, which are the through-space distances between amino acid residues that are separated by 20–56 residues, span the range of 46.0–54.2 Å (Table 1).

We have also measured $\langle E_{app} \rangle$ for each of our constructs unfolded in 50% TFE. Under these conditions, energy transfer is uniformly more

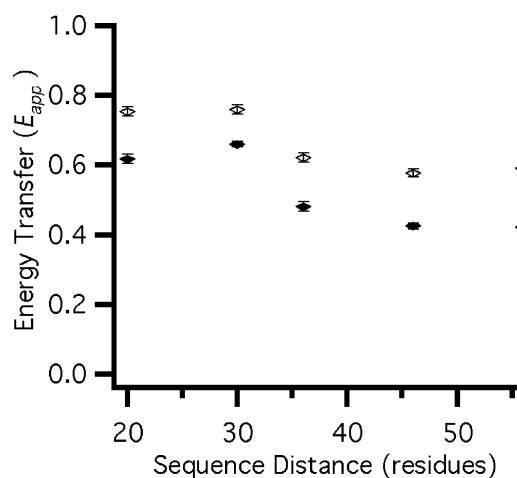


Figure 4. Observed mean transfer efficiency, $\langle E_{app} \rangle$, in 4 M GuHCl (filled symbols) and 50% TFE (open symbols). The relatively large difference in Forster radii in the two solvents produces dramatically different $\langle E_{app} \rangle$.

Table 1. Measured energy transfer efficiencies and calculated distances

Construct	Sequence distance (residues)	Energy transfer efficiency		Distance (Å)	
		GuHCl	TFE	GuHCl	TFE
S115C/S135C	20	0.62 ± 0.01	0.76 ± 0.01	47.4 ± 0.5	47.0 ± 0.6
T85C/S115C	30	0.66 ± 0.01	0.75 ± 0.01	46.0 ± 0.3	46.7 ± 0.6
K105C/V141C	36	0.48 ± 0.01	0.62 ± 0.01	52.1 ± 0.5	52.1 ± 0.5
A95C/V141C	46	0.43 ± 0.01	0.58 ± 0.01	53.4 ± 0.3	53.8 ± 0.4
T85C/V141C	56	0.42 ± 0.01	0.59 ± 0.01	54.2 ± 0.3	53.3 ± 0.3

efficient than that observed in GuHCl (Figure 4). For the dyes employed, however, R_0 is greater in TFE than it is in GuHCl. Using the calculated R_0 in 50% TFE (see below), we can convert the observed $\langle E_{app} \rangle$ into effective distances, R_{eff} , and find that the inter-residue distances in TFE are effectively indistinguishable from those observed in GuHCl (Table 1).

Forster radii

Forster radii, R_0 , are dependent on the spectral overlap of the dyes, $J(\lambda)$, an orientation factor, κ^2 , the quantum yield of the donor, ϕ_D , and refractive index, n , via the equation:²⁷

$$R_0 = 9.79 \times 10^3 (k^2 n^{-4} \phi_D J(\lambda))^{1/6} \text{ \AA} \quad (4)$$

Because both absorption and emission spectra vary with solvent, we collected donor emission and acceptor absorption spectra for the dyes conjugated to the S115C single-cysteine mutant in 4 M GuHCl and 50% TFE to calculate $J(\lambda)$ (Figure 5). We also measured the fluorescence lifetime of a protein-attached Alexa-488 (data not shown) and used this and previous measurements²⁸ to estimate $\phi_D = 0.63$ in 4 M GuHCl and 0.76 in 50% TFE. Rotational freedom of the dyes around their C₅ linkers has generally (and reasonably) been assumed to produce an averaged orientation factor, κ^2 , of 2/3.¹⁹ In order to test this, we have performed steady-state anisotropy measurements on all five of our heterolabeled constructs, and on both the donor and acceptor dyes attached to a single cysteine mutant (S115C). The anisotropy of all seven of these constructs is <0.15 under denaturing conditions (data not shown). Given that rigid, randomly oriented dyes would produce a κ^2 of 0.476, and that anisotropies even as high as 0.4 reduce to errors in R_{eff} of less than 10%,²⁷ the low anisotropies we observe support strongly our assumption that $\kappa^2 = 2/3$. The large-scale conformational transitions observed in the 1 ns molecular dynamics simulation (Figure 6) also supports the assertion that κ^2 converges on its average value during the excited state lifetime of the donor dye.¹⁹ From the above parameters, we calculate that R_0 is 51.4 Å in 4 M GuHCl and 56.7 Å in 50% TFE, in good agreement with previous reports.^{24,25}

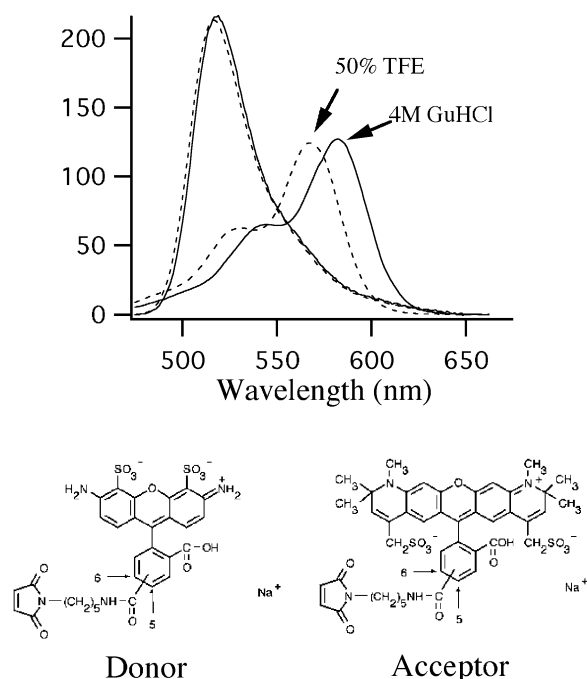


Figure 5. Forster radii can be derived from the spectral overlap of emission and absorption of donor and acceptor dyes. Shown are the absorption (left peaks) and emission (right peaks) spectra of dye-labeled proteins in 4 M GuHCl or 50% TFE. The emission spectra were normalized to unit area and the absorption spectra were normalized to the known extinction coefficient of Alexa 594 in phosphate buffer. The overlap integrals, $J(\lambda)$, were calculated from these spectra and used to determine the Forster radii under the conditions employed here. Shown below are the structures of the donor and acceptor dyes (Alexa 488 and Alexa 594, respectively).

Dye-linker offset

The observed dye-to-dye distance will differ from the true residue-to-residue distance because the dyes employed are of finite size and are attached to the polypeptide by long, flexible linkers. In order to estimate this offset, we have performed atomistic simulations of a solvated Alexa 488–cysteine conjugate. The RMS distance between the center of the fluorophore and the C α averaged over the 1 ns trajectory is 14.7 Å. Given, however, that the linker is flexible (as shown by large-scale transitions in the simulation, Figure 6), it is more accurate to view this offset in terms of the additional sequence distance, ζ , rather than as an absolute distance offset. Flory scaling laws predict that 14.7 Å corresponds to a six residue, random coil polypeptide

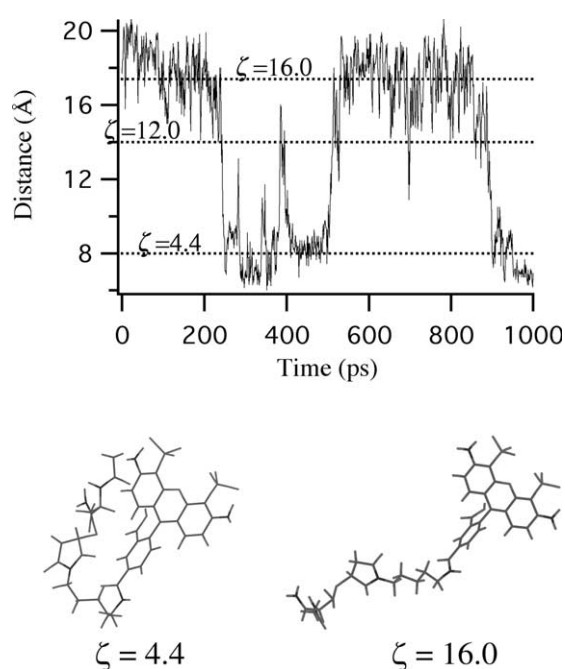


Figure 6. In order to estimate the offset produced by the finite size of the dyes and their linkers, we have performed a molecular dynamics simulation of a solvated Alexa 488–cysteine conjugate. The distance between the center of the dye and the cysteine C α is indicated. While the ensemble mean indicates that the two dye-linker pairs add the equivalent of 12 residues to the random coil polymer length, the ensemble appears to be composed of two families of structures with offsets, ζ , of 4.4 residues and 16.0 residues. The former family, however, appears to arise due to hydrophobic interactions between the dye and its linker that would presumably be abolished under denaturing conditions.

(see description below), and in turn to $\zeta=12$ residues for the two dye-linker pairs. Detailed inspection of our simulation results, however, suggests that this average obscures a highly bimodal distribution (Figure 6). The two structural families produce RMS distances of 8.1 Å and 17.4 Å, corresponding to offsets of $\zeta=4.4$ residues and 16.0 residues, respectively. Given the approximations inherent in these simulations, one or the other of these families may be better representative of the true offset. This is particularly true because the shorter family appears to occur due to hydrophobic interactions between the dye and its linker that would presumably be abolished in denaturant.

How does R_{eff} relate to N for a random coil polymer?

Energy transfer efficiency scales with an ensemble weighted average of the donor–acceptor distance. Thus, from $\langle E_{\text{app}} \rangle$ we calculate an effective, averaged distance, R_{eff} . In order to understand how R_{eff} would be expected to scale with N for a

random coil ensemble we must explore the properties of R_{eff} in detail.

At short sequence separations, chain stiffness starts to interfere with the ability of a chain to loop back onto itself and thus the relationship between N and R_{eff} will be complicated unless N is significantly longer than the persistence length, l_p . Such effects, however, have not been observed previously in denatured proteins and peptides, even at lengths as short as 16 amino acid residues,⁹ suggesting we are in the $N > l_p$ regime.

If we are in the $N > l_p$ regime, what is the relationship we should expect between R_{eff} and N ? The probability distribution of the vector \mathbf{R} connecting two ends of a random coil polymer chain that has $N \gg 1$ links has the scaling form:^{29,30}

$$p(\mathbf{R}) = \frac{1}{X^3} f(R/X) \quad (5)$$

where f is a universal function and X is related to the mean square end-to-end distance

$$X^2 = \langle R^2 \rangle / 3 \quad (6)$$

The latter satisfies the scaling law:

$$\langle R^2 \rangle \propto N^{2\nu} \quad (7)$$

where $\nu \approx 3/5$ is Flory's exponent for chains with excluded volume in a good solvent.

$\langle E \rangle$ is then given by:

$$\begin{aligned} \langle E \rangle &= 1 - \left\langle \frac{R^6}{R^6 + R_0^6} \right\rangle \\ &= 1 - X^{-3} \int_0^\infty 4\pi R^2 f(R/X) \frac{R^6}{R^6 + R_0^6} dR \\ &= 1 - \Phi(R_0/X) \end{aligned} \quad (8)$$

where the dimensionless function $\Phi(s)$ is defined by the equation:

$$\Phi(s) = \int_0^\infty 4\pi y^2 f(y) \frac{y^6}{y^6 + s^6} dy \quad (9)$$

where $s = (R_0/X)$. The average FRET efficiency corresponds to R_{eff} via equation (3), so that:

$$R_{\text{eff}} = R_0 \left[\frac{1}{\Phi(R_0/X)} - 1 \right]^{-1/6} \quad (10)$$

(While the dimensions of a chain, as defined by its end-to-end distance, scales according to Flory's power law, equation (7), R_{eff} generally does not, because equation (10) involves two different length scales, X and R_0 .) To compute R_{eff} , we need to know $f(y)$. Here, we use the interpolation function of the form:³⁰

$$\begin{aligned} f(y) &= f_1 y^\theta \exp[-Dy^\delta], \quad \theta = 0.275, \\ \delta &= (1 - \nu)^{-1} \end{aligned} \quad (11)$$

that satisfies the known scaling laws,^{29,30} both for $y \rightarrow 0$ and $y \rightarrow \infty$; The coefficients $D=0.335$ and $f_1=0.0495$ are determined from the condition that $f(y)$ is normalized and satisfies equation (6).

Note that equation (11) jointly with equation (5) describes the dependence of the probability distribution for the distance between the ends of a polymer chain. Our donor and acceptor, however, are attached *via* relatively long linkers to side-chains within the polymer. And, while scaling laws of the type of equation (11) have been found,³⁰ for monomer pairs situated within an infinitely long chain and for a chain end and another monomer belonging to a semi-infinite chain (data not shown), the sequence separations explored here are comparable to the chain length, so neither limit is applicable. By experimenting with 3D cubic lattice walks of $N \leq 100$, however, we find that equation (11) holds in the experimentally relevant regime (data not shown). Further, in the experimental range of N , the numerical result for R_{eff} is rather insensitive to the exact values of θ and δ : over the experimentally appropriate range of θ and δ , the random coil distances derived from the semi-infinite and end-to-end models differ by no more than ~ 1 Å (data not shown).

Asymptotic scaling laws can be obtained for R_{eff} in two limits. In the limit when R_{eff} is short relative to the Forster radius ($s=R_0/X \gg 1$) then:

$$\langle E \rangle \approx 1 - \left\langle \frac{R^6}{R_0^6} \right\rangle \approx 1 - \frac{R_{\text{eff}}^6}{R_0^6} \quad (12)$$

which gives:

$$R_{\text{eff}} \approx \langle R^6 \rangle^{1/6} \quad (13)$$

Using equation (5), we have:

$$\langle R^6 \rangle = X^6 \int_0^\infty dy 4\pi y^8 f(y) \quad (14a)$$

$$\langle R^6 \rangle^{1/6} = A \langle R^2 \rangle^{1/2};$$

$$A = \left[\int_0^\infty dy 4\pi y^8 f(y) \right]^{1/6} / \sqrt{3} \quad (14b)$$

From equation (11) we estimate that $A \approx 1.2$. We have also computed A directly by performing Monte Carlo simulations of random-flight, hard-core repulsion polyaniline chains.^{31–33} In these simulations, the ratio $A(N) = \langle R^6 \rangle^{1/6} / \langle R^2 \rangle^{1/2}$ increased from 1.10 to 1.176 as N varied from 15 to 130 (Figure 7). Extrapolation of our Monte Carlo data to $N \rightarrow \infty$ gives $A = A(\infty) \approx 1.19$. This is within error of the extrapolated value $A(\infty)$ that we estimate for 3D cubic lattice self-avoiding walks of $20 \leq N \leq 100$ (Figure 7). Equations (12) and (14a) indicate that R_{eff} obeys the Flory scaling law in this limit and at short distances the polymer will scale as $R_{\text{eff}} = \langle R^6 \rangle^{1/6}$.

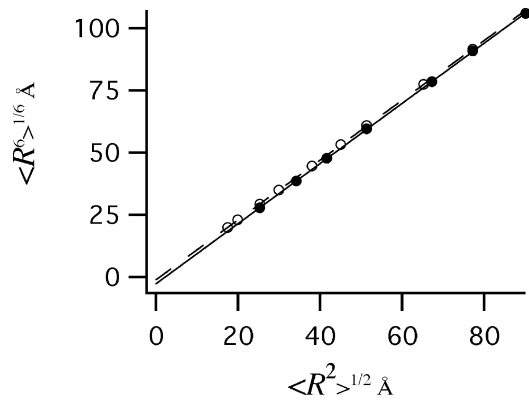


Figure 7. Simulations of both lattice polymers (open symbols) and a fully atomistic model of polyaniline (filled symbols) demonstrate that the relationship $\langle R^6 \rangle^{1/6} = 1.2 \langle R^2 \rangle^{1/2}$ holds over a wide range of polymer lengths and polymer types.

When Flory scaling is obeyed, the expected value of R_{eff} for a random coil can be derived from scattering measurements. Extensive studies demonstrate that the RMS radius of gyration, $\langle R_g^2 \rangle^{1/2}$, is equal to $l_0 N^{0.588}$ where $l_0 = 2.05$ Å.⁹ And both renormalization group studies and simulations of self-avoiding walks³⁰ indicate that, for the excluded volume polymers, the relationship between end-to-end distance, R , and R_g is given by $\langle R^2 \rangle = 6.316 \langle R_g^2 \rangle$. Using this relationship we have:

$$\langle R^2 \rangle^{1/2} = 2.51 \langle R_g^2 \rangle^{1/2} = 5.15(N + \zeta)^{0.588} \text{ Å} \quad (15)$$

where ζ is the offset due to the finite dye-linker size determined from molecular dynamics of the donor molecule, as determined above. Correcting for the difference between $\langle R^2 \rangle^{1/2}$ and $\langle R^6 \rangle^{1/6}$, we have:

$$R_{\text{eff}} = \langle R^6 \rangle^{1/6} = A \langle R^2 \rangle^{1/2} = 6.03(N + \zeta)^{0.588} \text{ Å} \quad (16)$$

for a random coil polypeptide when R_{eff} is short relative to the Forster radius.

In the limit when R_{eff} is long relative to the Forster radius ($s=R_0/X \gg 1$) equations (9)–(11) result in a different asymptotic expression:

$$R_{\text{eff}} \approx 1.2 R_0^{\frac{1-\theta}{6}} X^{\frac{1+\theta}{6}} \text{ Å} \quad (17)$$

In this limit, therefore, we have:

$$R_{\text{eff}} \propto N^{\nu(\frac{1+\theta}{6})} \approx N^{0.321} \quad (18a)$$

$$\begin{aligned} R_{\text{eff}} &= 0.889 R_0^{0.454} (\langle R^2 \rangle^{1/2})^{0.546} \\ &= 13.3(N + \zeta)^{0.321} \text{ Å} \end{aligned} \quad (18b)$$

which differs significantly from Flory scaling. Equation (18b) is the applicable scaling law in the limit of large N (i.e. $N \rightarrow \infty$). (We note for comparison that, $R_{\text{eff}} \propto N^{1/4}$ for a Gaussian polymer.³⁴) Since the experimental values of N fall

between these two limits, we have computed R_{eff} numerically using equations (6), (9), (10), and (11).

Discussion

Our ability to convert energy transfer efficiency to absolute residue-to-residue distances depends on the validity of several key assumptions: (1) that we are able to measure transfer efficiency accurately; (2) that the orientation factor, κ^2 , has converged on its average value for all constructs; (3) that the calculated Forster radii are correct; (4) that our estimate for the offset, ζ , produced by the dyes and their linkers is accurate; and (5) that the dyes do not perturb the structure of the denatured state. To the extent that these critical assumptions hold, and thus that we can measure residue-to-residue distances accurately, we find that the distances across GuHCl-denatured FynSH3 approximate the end-to-end dimensions expected for a random coil chain (Figure 8). Such behavior has long been considered a hallmark of random coil ensembles,³⁵ and is consistent with the results of extensive SAXS studies suggesting that proteins adopt a globally random coil conformation under highly denaturing conditions.^{8,9,36}

While the mean residue-to-residue dimensions of the GuHCl-denatured protein approximate those expected for a random coil, we nevertheless observe small, but experimentally significant deviations from random coil behavior, with some inter-residue distances deviating by up to $\sim 10\%$ from expected random coil values. Some portion of the observed deviation may arise because our estimate of how

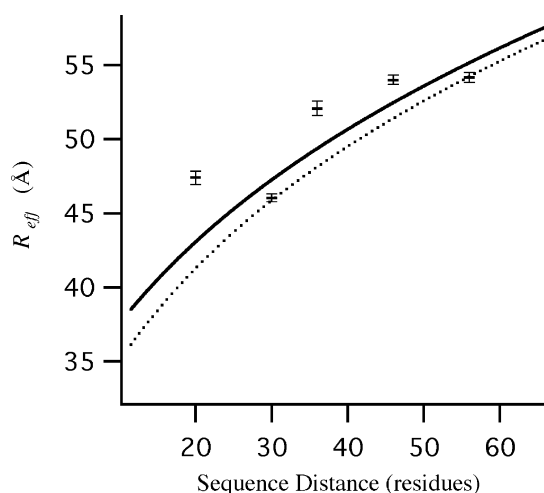


Figure 8. Residue-to-residue distances in the GuHCl-unfolded state roughly approximate those expected for a random coil (the upper and lower lines correspond to offsets of $\zeta=12$ residues and 16 residues, respectively). The random coil prediction was derived entirely from theory, simulations (to estimate ζ) and prior, independent small-angle scattering studies (to determine the pre-factor) and is not fitted to the observed distances.

R_{eff} scales with N for a random coil assumes a semi-infinite chain (see the discussion of equation (11)). And, while similar scaling laws have been found for monomer pairs situated within an infinite chain,³⁰ the constructs characterized here do not fit either of these regimes precisely. Over the range of sequence separations we have characterized, however, estimates produced by these two models vary by no more than ~ 1 Å, suggesting that other factors dominate the observed deviations. Alternatively, Schuler *et al.*²⁵ have suggested that, in addition to the above described potential caveats, two additional mechanisms can produce errors in our estimates of R_{eff} : (1) the exaggerated distances measured at short sequence separation may be due to the failure of the point-dipole approximation (i.e. are due to the finite size of the dyes); and (2) our polymer scaling model is static and does not account for chain diffusion over the lifetime of the donor. These systematic effects, however, would be expected to produce deviations from random coil behavior that vary monotonically with polymer length rather than the non-monotonic deviations we observe. Lastly, for short sequence separations (i.e. when N approaches the persistence length) deviations from random coil behavior may occur due to chain stiffness. SAXS studies⁹ suggest, however, that the persistence length of an unfolded polypeptide is significantly shorter than the shortest sequence-separation studied here and thus such effects seem unlikely to account for the observed deviations. While it is difficult to categorically rule out sources of systematic experimental error that may vary from construct to construct, the observed deviations may represent real excursions from random coil behavior and thus may reflect residual structure in the chemically denatured state of this protein.

If the observed deviations between expected and observed R_{eff} represent real excursions from random coil behavior, they may be consistent with both spectroscopic studies of highly denatured proteins and with recent simulations demonstrating that a protein could exhibit random coil dimensions even while retaining very significant sequence-local structure.^{16,36} Of specific relevance to our claims, Zhang & Forman-Kay³⁷ have reported $i, i+2$ and $i, i+3$ nuclear Overhauser effect (NOE) connectivities in a homologous SH3 domain unfolded in 2 M GuHCl, suggesting that under these conditions the protein populates some sequence-local structure, which presumably would lead to deviations from random coil scaling similar to those observed here. Our conclusions may be consistent also with recent claims that the denatured state adopts a grossly native-like topology.³⁸ Nevertheless, we observe no significant correlation between distances across the unfolded and native states (data not shown), an observation that is consistent with the results of other recent studies that attribute residual dipole couplings to the presence of transient local structure, such as sequences preferentially populating the β -strand

and nonnative polyproline(II) helix minima on the Ramachandran energy surface.¹⁰

In contrast to the GuHCl-denatured state, which produces an effectively random coil backbone conformation,⁸ TFE induces an unfolded state that exhibits strong circular dichroism at 220 nm, a spectroscopic feature characteristic of helices (Figure 2). While even such a highly helical state should behave as a random coil (of short, helical elements), the mean dimensions of this random coil might be expected to differ significantly from those of a random coil comprised of locally unstructured elements. Nevertheless, we find that the five residue-to-residue distances measured in TFE are within error of those observed in GuHCl (Figure 9). This perhaps surprising result is in keeping, however, with previously reported scattering experiments using TFE, urea and other helix-inducing and coil-inducing solvents, which indicates that the mean dimensions of the two unfolded states do not differ significantly.⁸

The results reported here suggest a means of reconciling the apparently conflicting results of scattering experiments, which indicate a random coil unfolded state,⁹ and a variety of spectroscopic probes suggesting that the unfolded state contains significant residual structure.¹⁻⁷ That is, when averaged over five constructs the mean dimensions of the GuHCl-unfolded protein approximate closely those expected for a random coil ensemble. In contrast, individual measurements suggest that this globally average behavior may hide real deviations from random coil behavior. The results presented here suggest also that the similarity in the mean,

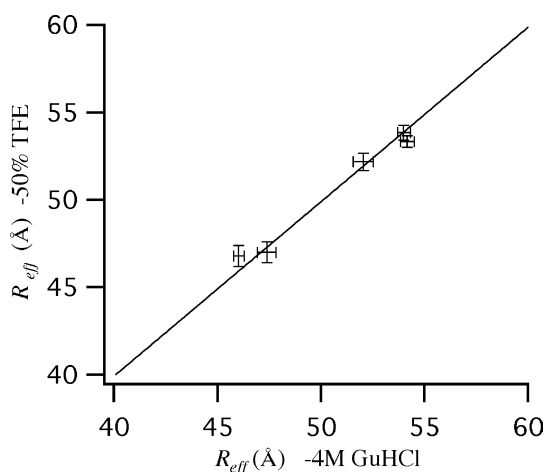


Figure 9. Whereas the energy transfer observed in 50% TFE differs significantly from that observed for the same constructs in 4 M GuHCl (Figure 4), this discrepancy arises entirely due to solvent effects on the R_0 . When this is taken into account, the five distances measured across the TFE-unfolded protein are indistinguishable from those measured in GuHCl (fitted slope = 0.998 ± 0.005). This occurs despite the highly helical unfolded state induced by TFE, suggesting that the gross structure of the denatured state is independent of the chemistry of the cosolvent.

global dimensions of proteins unfolded in TFE and GuHCl (as measured by scattering⁸) may extend to the finer details of long-range geometry of these unfolded states. This, in turn, suggests that the gross structure of the denatured state is, perhaps surprisingly, independent of the chemistry of the cosolvent and of the nature of any residual, sequence-local structure.

Materials and Methods

Sample preparation

His-tagged, double cysteine mutants (Figure 1) were generated and confirmed by sequencing or mass spectroscopy. The proteins were expressed (pET vector in *Escherichia coli*) and purified by affinity chromatography (Qiagen) and reverse-phase HPLC (C₄, HP 1100, Waters).

The proteins were labeled simultaneously with Alexa Fluor 488 C₅ maleimide and Alexa Fluor 594 C₅ maleimide (Figure 5) (Molecular Probes) in 50 mM NH₄HCO₃, 3.6 M GuHCl (Pierce), and a 20-fold excess of tris(2-carboxyethyl)phosphine HCl (Fluka). The dyes and protein were mixed in a 10:10:1 molar ratio. Excess dye was removed using a Ni-Agarose column. This procedure creates heterolabeled, homolabeled and mono-labeled molecules. For ensemble measurements, such heterogeneity would lower the apparent transfer efficiency. Single-molecule measurements, however, allow us to discard data arising from mislabeled molecules. Nevertheless, to minimize background arising from "donor-only" labeled material, additional HPLC purification was performed (YMC Pack C₄ Protein RP, Waters). HPLC-purified proteins were lyophilized, taken up in Dulbecco's phosphate-buffered saline and stored at -70°C .

Single-molecule spectroscopy

Dye-labeled protein was diluted to ~ 0.1 nM in 4 M GuHCl (Fluka) or 50% (v/v) 2,2,2 trifluoroethanol (TFE) (FisherBiotech), and 0.01% (v/v) Tween-20 in Dulbecco's phosphate-buffered saline. These solutions were pipetted onto a cover-slip and positioned above a Leitz 100 \times magnification 1.2NA water-immersion objective; 200 μW of 496 nm light from a mode-locked argon ion laser (Spectra Physics model 2080) was focused to a near diffraction limited spot ~ 25 μm into the sample. The 50% TFE samples were placed between two cover-slips with a silicone spacer to minimize evaporation. Excitation light was reflected into the objective by a 505DRLP dichroic mirror (Omega Optical), and collected fluorescence was passed back through this dichroic, spatially (100 μm pinhole) and spectrally (560DRLP dichroic) filtered, and focused onto the detectors. Detection employed single-photon counting avalanche photodiodes (Perkin-Elmer Optoelectronics) masked using band-pass filters (530df30 and 630df30, Omega Optical) and a SPC630 photon counting card and router (Becker&Hickl). Data collection was for ~ 1 h, during which tens of thousands of single-molecule transits were observed.

We extracted single-molecule transits by binning the raw data into 1 ms intervals and identifying bursts (arising due to molecules diffusing across the excitation volume) as events exceeding a threshold number of photoelectrons. We find that E_{app} is independent of

threshold size over the range from 25–50 photoelectrons, and thus we adopted the former value as our burst threshold.

Maximum likelihood estimation

E_{app} is the probability that a photoelectron will be detected in the acceptor channel (equation (1)) and its mean is determined from the probability distribution. We must keep in mind that two underlying distributions are observed in this analysis: one arising due to molecules lacking an acceptor dye, and a second from properly labeled molecules (equation (19a)). We define the probability that a photoelectron will be in the acceptor channel as p_0 if that photon originates from a molecule lacking an acceptor and p_1 if that photon arises from a properly tagged molecule. We are interested in estimating p_1 , which is equivalent to $\langle E_{\text{app}} \rangle$ for properly labeled molecules. For a particular single-molecule transit, we observe a total n_i photoelectrons at time-point i . Then the number of detected acceptor photoelectrons is Bernoulli(n_i, p_i) (in an ideal case where there is no background or acceptor photobleaching and shot noise dominates), where p_i is either p_0 or p_1 (equation (19b)). Assume that the probability of observing photoelectrons from a molecule without an acceptor dye at any time, p , is independent of the total number of photoelectrons observed. Let X be the number of acceptor photoelectrons. Then:

$$P(X = x_i | n_i) = \sum_{p_i} P(X = x_i | n_i, p_i) \times P(p_i | n_i) \quad (19a)$$

$$= P(X = x_i | n_i, p_0) \times p + P(X = x_i | n_i, p_1) \times (1 - p) \quad (19b)$$

$$= \binom{n_i}{x_i} p_0^{x_i} (1 - p_0)^{n_i - x_i} \times p + \binom{n_i}{x_i} p_1^{x_i} (1 - p_1)^{n_i - x_i} \times (1 - p) \quad (19c)$$

and the likelihood is:

$$L(p, p_0, p_1) =$$

$$\prod_{i=1}^N \left[\binom{n_i}{x_i} p_0^{x_i} (1 - p_0)^{n_i - x_i} \times p + \binom{n_i}{x_i} p_1^{x_i} (1 - p_1)^{n_i - x_i} \times (1 - p) \right] \quad (20)$$

where N is the number of bursts with $n_i \geq 25$, and all bursts are assumed independent. For these observations, given the number of photoelectrons in the acceptor channel (x_i) and the total number of photoelectrons (n_i), we use maximum likelihood to obtain parameter estimates for the probabilities p , p_0 , and p_1 . The results are demonstrated by plotting the number of photoelectrons in the acceptor channel, x_i , against total burst size, n_i , and fit with maximum likelihood estimation to give the fit lines $p_i \cdot n_i$ (Figure 3, top; p_0 heavy line and p_1 light line; additional data are provided as Supplementary Data). These data can be presented as a histogram of E_{app} probabilities (Figure 3, bottom). Analyzing the data in this fashion makes the correct assumptions about distribution for the number of acceptor photons observed given a burst size, and therefore inference *via* maximum likelihood yields correct estimates for the mean and standard error of E_{app} .

Standard errors for the estimates of p , p_0 , and p_1 in each experiment were obtained by calculating the square-root of the diagonal elements in the inverse of the information

matrix. The information matrix is the negative of the Hessian, defined as the matrix of second derivatives of the log likelihood.³⁹ The standard errors for p_1 reflect the uncertainty about the estimate for each p_1 , measured in separate experiments, and were therefore used as inverse weights to obtain the estimates for the mean responses for the constructs. It is important to realize that there exists a second source of variability: because of the variability in the experimental conditions, the “true” underlying parameters between the experiments for the same construct also differ. To take this into account, we calculated a pooled variance from two to five experiments per construct, assuming equal between-experimental variability for all constructs. This pooled variance was used to calculate the standard error for the estimate of the mean of each construct.

Systematic influences

Photon counts between bursts suggest that about one to three photoelectrons per burst arise due to laser scatter and detector dark counts (data not shown). Given our 25 photon threshold, and given that $\langle E_{\text{app}} \rangle$ is ~ 0.5 for all constructs, this background is negligible. $\langle E_{\text{app}} \rangle$ does not change systematically with burst size, supporting this assertion (data not shown).

Fluorescence correlation spectroscopy indicates that at the concentrations employed, the probe volume is occupied $< 10\%$ of the time (data not shown), and two molecule events thus occur $< 1\%$ of the time. Because two molecules are more likely than one to produce > 25 photons, however, we estimate that up to 5% of the bursts over this threshold are from two-molecule events.⁴⁰ If both molecules are heterolabeled, $\langle E_{\text{app}} \rangle$ is not affected. In contrast, if one of the molecules lacks an acceptor, E_{app} will be one-third to one-half of the true value. Given the fact that the majority of the molecules are properly labeled, however, the peak arising from improper two-molecule events is much less than 5% and will not affect the reported results significantly. Consistent with this, the exclusion of bursts > 100 photoelectrons (which are more likely to reflect two-molecule events) does not alter our results significantly (data not shown).

Dye geometry

The C_5 dye linkers increase the mean donor–acceptor distance. In order to estimate the size of this effect, we simulated Alexa 488 C_5 maleimide-cysteine in water using the Amber 7.0 suite of programs[†]. The dye was constructed using Hyperchem 7.0 (Hypercube Inc., 2002), minimized, and solvated in a box containing ~ 500 TIP3P water molecules. A series of equilibration steps were then applied to allow water molecules to relax around the solute at a density of 1 kg/l. The system was then annealed by minimization, followed by 30 ps of dynamics at 400 K (two-step heating cycle 100–400 K, 300–400 K for 2 ps each and 26 ps of 400 K dynamics), cooling to 50 K in 1 ps intervals, and finally 30,000 steps of additional minimization. Starting from this structure, we performed a 1 ns, 300 K molecular dynamics simulation and measured the dye- C^{α} distance each picosecond (Figure 6).

[†] <http://amber.scripps.edu>.

Polymer chain simulations

Random flight, hard shell repulsion polyalanine chain ensembles were generated using a Monte Carlo procedure described elsewhere.^{30,31} Random walks on a 3D cubic lattice were generated by the Rosenbluth sampling method.⁴¹ The results of these simulations are shown in Figure 7.

Acknowledgements

The authors thank Richard Keller and Thomas Louis for helpful comments and discussions. This work was supported through grants from the Robert A. Welch foundation, the ACS PRF and the NSF (CAREER award) to D.E.M., a faculty innovation award from John Hopkins University to I.R., the Los Alamos National Laboratory LDRD program to J.H.W. and P.M.G., and CULAR and NIH (RO1GM62868-01A2) funds to K.W.P. S.L.B. was supported by the NSF (grant to M.T. Bowers).

Supplementary Data

Supplementary data associated with this article can be found at [10.1016/j.jmb.2005.07.015](https://doi.org/10.1016/j.jmb.2005.07.015)

References

- Kazmirski, S. L., Wong, K. B., Freund, S. M. V., Tan, Y. J., Fersht, A. R. & Daggett, V. (2001). Protein folding from a highly disordered denatured state: the folding pathway of chymotrypsin inhibitor 2 at atomic resolution. *Proc. Natl Acad. Sci. USA*, **98**, 4349–4354.
- Garcia, P., Serrano, L., Durand, D., Rico, M. & Bruix, M. (2001). NMR and SAXS characterization of the denatured state of the chemotactic protein CheY: implications for protein folding initiation. *Protein Sci.* **10**, 1100–1112.
- Hodsdon, M. E. & Frieden, C. (2001). Intestinal fatty acid binding protein: the folding mechanism as determined by NMR studies. *Biochemistry*, **40**, 732–742.
- Klein-Seetharaman, J., Oikawa, M., Grimshaw, S. B., Wirmer, J., Duchardt, E., Ueda, T. *et al.* (2002). Long-range interactions within a nonnative protein. *Science*, **295**, 1719–1722.
- Baldwin, R. L. (2002). Protein folding—making a network of hydrophobic clusters. *Science*, **295**, 1657–1658.
- Shortle, D. & Ackerman, M. S. (2001). Persistence of native-like topology in a denatured protein in 8 M urea. *Science*, **293**, 487–489.
- Shortle, D. (1996). The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB J.* **10**, 27–34.
- Millet, I., Doniach, S. & Plaxco, K. W. (2002). Toward a taxonomy of the denatured state: small angle scattering studies of unfolded proteins. *Advan. Protein Chem.* **62**, 321–325.
- Kohn, J. E., Millett, I. S., Jacob, J., Dillon, T. M., Cingel, N., Dothager, R. S. *et al.* (2004). Random coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl Acad. Sci. USA*, **101**, 12491–12496.
- Mohana-Borges, R., Goto, N. K., Kroon, G. J. A., Dyson, H. J. & Wright, P. E. (2004). Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings. *J. Mol. Biol.* **340**, 1131–1142.
- Neri, D., Billeter, M., Wider, G. & Wuthrich, K. (1992). NMR determination of residual structure in a urea-denatured protein, the 434-repressor. *Science*, **257**, 1559–1563.
- Tafer, H., Hiller, S., Hilty, C., Fernandez, C. & Wuthrich, K. (2004). Nonrandom structure in the urea-unfolded *Escherichia coli* outer membrane protein X (OmpX). *Biochemistry*, **43**, 860–869.
- Bhavesh, N. S., Panchal, S. C., Mittal, R. & Hosur, R. V. (2001). NMR identification of local structural preferences in HIV-1 protease tethered heterodimer in 6 M guanidine hydrochloride. *FEBS Letters*, **509**, 218–224.
- Bhavesh, N. S., Juneja, J., Udgaonkar, J. B. & Hosur, R. V. (2004). Native and nonnative conformational preferences in the urea-unfolded state of barstar. *Protein Sci.* **13**, 3085–3091.
- Yi, Q., Scalley-Kim, M. L., Alm, E. J. & Baker, D. (2000). NMR characterization of residual structure in the denatured state of protein L. *J. Mol. Biol.* **299**, 1341–1351.
- Fitzkee, N. C. & Rose, G. D. (2004). Reassessing random coil statistics in unfolded proteins. *Proc. Natl Acad. Sci. USA*, **101**, 12497–12502.
- Plaxco, K. W., Gujjarro, J. I., Morton, C. J., Pitkeathly, M., Campbell, I. D. & Dobson, C. M. (1998). The folding kinetics and thermodynamics of the Fyn-SH3 domain. *Biochemistry*, **37**, 2529–2537.
- Ha, T., Ting, A. Y., Caldwell, W. B., Deniz, A. A., Chemla, D. S., Schultz, P. G. & Weiss, S. (1999). Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism. *Proc. Natl Acad. Sci. USA*, **96**, 893–898.
- Deniz, A. A., Laurence, T. A., Beligere, G. S., Dahan, M., Martin, A. B., Chemla, D. S. *et al.* (2000). Single-molecule protein folding: diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *Proc. Natl Acad. Sci. USA*, **97**, 5179–5184.
- Borsch, M., Diez, M., Zimmermann, B., Reuter, R. & Graber, P. (2002). Stepwise rotation of the gamma-subunit of EF0F1-ATP synthase observed by intramolecular single-molecule fluorescence resonance energy transfer. *FEBS Letters*, **527**, 147–152.
- Katiliene, Z., Katilius, E. & Woodbury, N. W. (2003). Single molecule detection of DNA looping by NgoMIV restriction endonuclease. *Biophys. J.* **84**, 4053–4061.
- Slaughter, B. D., Allen, M. W., Unruh, J. R., Urbauer, R. J. B. & Johnson, C. K. (2004). Single-molecule resonance energy transfer and fluorescence correlation spectroscopy of calmodulin in solution. *J. Phys. Chem. B*, **108**, 10388–10397.
- Xie, Z., Srividya, N., Sosnick, T. R., Pan, T. & Scherer, N. F. (2004). Single-molecule studies highlight conformational heterogeneity in the early folding steps of a large ribozyme. *Proc. Natl Acad. Sci. USA*, **101**, 534–539.
- Schuler, B., Lipman, E. A. & Eaton, W. A. (2002). Proving the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, **419**, 743–747.

25. Schuller, B., Lipman, E. A., Steinback, P. J., Kumke, M. & Eaton, W. A. (2005). Polyproline and the "spectroscopic ruler" revisited with single-molecule fluorescence. *Proc. Natl Acad. Sci. USA*, **102**, 2754–2759.
26. Rhoades, E., Gussakovskiy, E. & Haran, G. (2003). Watching proteins fold one molecule at a time. *Proc. Natl Acad. Sci. USA*, **100**, 3197–3202.
27. Lakowicz, J. R. (1999). *Principles of Fluorescence Spectroscopy* (2nd edit.), Kluwer Academic/Plenum Publishing, New York.
28. Rothwell, P. J., Berger, S., Kensch, O., Felekyan, S., Antonik, M., Wohrl, B. M. *et al.* (2003). Multiparameter single-molecule fluorescence spectroscopy reveals heterogeneity of HIV-1 reverse transcriptase:primer/template complexes. *Proc. Natl Acad. Sci. USA*, **100**, 1655–1660.
29. De Gennes, P. G. (1979). *Scaling Concepts in Polymer Physics*, Cornell University Press, Ithaca.
30. Des Cloizeaux, J. & Jannink, G. (1990). *Polymers in Solution: Their Modeling and Structure*, Clarendon Press, Oxford.
31. Wang, Z. & Makarov, D. E. (2002). Rate of intramolecular contact formation in peptides: the loop length dependence. *J. Chem. Phys.* **117**, 4591–4593.
32. Makarov, D. E., Wang, Z., Thompson, J. & Hansma, H. G. (2002). On the interpretation of force extension curves of single protein molecules. *J. Chem. Phys.* **116**, 7760–7765.
33. Wang, Z. & Makarov, D. E. (2003). Nanosecond dynamics of single polypeptide molecules revealed by photoemission statistics of fluorescence resonance energy transfer: a theoretical study. *J. Phys. Chem. B*, **107**, 5617–5622.
34. Yang, S., Witkoskie, J. B. & Cao, J. (2002). Single-molecule dynamics of semiflexible Gaussian chains. *J. Chem. Phys.* **117**, 11010–11023.
35. Tanford, C., Kawahara, K. & Lapanje, S. (1966). Proteins in 6 M guanidine hydrochloride demonstration of random coil behavior. *J. Biol. Chem.* **241**, 1921–1923.
36. McCarney, E. R., Kohn, J. E. & Plaxco, K. W. (2005). Is there or isn't there? The case for (and against) residual structure in chemically denatured proteins. *Crit. Rev. Biochem. Mol. Biol.* In the press.
37. Forman-Kay, J. D. & Zhang, O. (1997). NMR studies of unfolded states of an SH3 domain in aqueous solution and denaturing conditions. *Biochemistry*, **36**, 3959–3970.
38. Ohnishi, S., Lee, A. L., Edgell, M. H. & Shortle, D. (2004). Direct demonstration of structural similarity between native and denatured Eglin C. *Biochemistry*, **43**, 4064–4070.
39. Lehmann, E. L. & Casella, G. (1998). *Theory of Point Estimation* (2nd edit.), Springer, New York.
40. Rigler, R. & Mets, U. (1992). Diffusion of single molecules through a Gaussian laser beam. *Laser Spec. Biomol.* **1921**, 239–248.
41. Rosenbluth, M. N. & Rosenbluth, A. W. (1955). Monte-Carlo calculations of the average extension of molecular chains. *J. Chem. Phys.* **23**, 356–359.

Edited by P. Wright

(Received 16 February 2005; received in revised form 1 July 2005; accepted 6 July 2005)
Available online 25 July 2005