SUPPORTING INFORMATION Hofmann *et al.*

Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single molecule spectroscopy

Preparation and labeling of proteins.

The cysteine containing variants of a destabilized variant of *human* cyclophilin A (W121F/C52/61/115/161S) (*h*CypA) were produced recombinantly in BL21DE3 as inclusion bodies (IBs). After cell disruption, 0.5 vol. of 60 mM EDTA, 6% Triton, 1.5 M NaCl were added and the raw extract was stirred at 4°C overnight. IBs were isolated by centrifugation at 48,200 g for 30 min at 10°C. The resulting IBs were washed with 0.1 M TrisHCl, 1 mM EDTA, pH 8 and resolubilized with 6 M GdmCl, 50 mM TrisHCl pH 7.5, 100 mM DTT. After centrifugation, the DTT was removed by desalting the resulting supernatant using a 26/60 desalting column (GE Healthcare) pre-equilibrated with 6 M GdmCl, 50 mM TrisHCl pH 8.0, 10 mM imidazole. The protein-containing fractions were immediately loaded on a HisTrap-column, and the His-tagged protein was eluted with a gradient from 0% to 100% 6 M GdmCl, 50 mM TrisHCl, 500 mM imidazole, pH 8. All hCypA-containing fractions were pooled and concentrated in the presence of 5 mM TCEP. The His-tag was cleaved by slowly adding 1-3 ml of hCypA to 40 ml of 50 mM TrisHCl, 0.5 M L-Arg, 1 mM TCEP containing 1.25 μM HRV C3-protease, pH 8. Since the variants of *h*CypA are highly destabilized compared to wt-*h*CypA, the variants aggregate during cleavage. After 2 hours, 3.5 M NH_4SO_4 were added to precipitate the protein. The suspension was centrifuged at 48,200 g for 1 hour at 10°C, and the pellet was dissolved in 2 ml of 6 M GdmCl, 50 mM TrisHCl, 10 mM imidazole, pH 8. The sample was then loaded on a HisTrap column (5 ml, GE Healthcare) with a high flow rate of 4 ml/min. The flow-trough contained 100-200 μM His-tag-free *h*CypA. To reduce the *h*CypA variants for labeling, 1 ml of the *h*CypA sample was incubated for 1 hour with 200 mM β-Mercaptoethanol and desalted afterwards using a HiTrap desalting column (5 ml, GE Healthcare) pre-equilibrated with 6 M GdmCl, 50 mM potassium phosphate pH 7.2. Immediately after elution, 0.5 equivalents of the donor fluorophore AlexaFluor 488 C5 maleimide (Invitrogen) was added. After 2 hours at room temperature, the reaction was stopped by the addition of 200 mM β-Mercaptoethanol. Unlabeled protein was separated from labeled protein using reversed phase chromatography (C18) with a gradient from aqueous 0.1 % trifluoroacetic acid (TFA) to 100% acetonitril without TFA. The pooled fractions were analyzed by mass spectrometry (ESI) and lyophilized. After resolubilization of the labeled *h*CypA variants in 6 M GdmCl, 50 mM potassium phosphate pH 7.2, a threefold excess of acceptor AlexaFluor 594 C5 maleimide (Invitrogen) was added. After 7 hours, 100 μM TCEP was added, and the doubly-labeled protein was separated from free dye using size-exclusion chromatography (6 M GdmCl, 50 mM potassium phosphate pH 7.2).

The spectrin domains R15 and R17 were expressed and purified as described by Scott *et al.* (1). For labeling of the spectrin domains, cysteine residues were introduced by site-

directed mutagenesis at positions 39 and 99 (R_{17}60 and R_{15}60) or 6 and 99 (R_{17}93 and R_{15}93). In R17, an endogenous cysteine at position 68 was exchanged to alanine to avoid multiple labeling. For labeling, a 1.3:1 molar excess of reduced protein was incubated with Alexa Fluor 488 maleimide (Invitrogen) at 4°C for ~10 hours. Un-reacted dye was removed by gel filtration (G25 desalting; GE Healthcare Biosciences AB, Uppsala, Sweden), and the protein was incubated with Alexa Fluor 594 maleimide at room temperature for ~2 hours. Doubly labeled protein was purified by ion-exchange chromatography (MonoQ HR 5/5; GE Healthcare Biosciences AB, Uppsala, Sweden).

The variants of the cold shock protein from *Thermotoga maritima* were produced and labeled as described in Soranno *et al.* (2). The purification and labeling of the intrinsically disordered proteins prothymosin α and the N-terminal domain of HIV integrase are described in Müller-Späth *et al.* (3).

Single-Molecule Fluorescence Spectroscopy.

Measurements were performed at 22 °C using either a custom-built confocal microscope as described previously (3, 4) or a Micro Time 200 confocal microscope equipped with a HydraHarp 400 counting module (Picoquant, Berlin, Germany). The donor dye was excited with a diode laser at 485 nm (dual mode: continuous wave and pulsed, LDH-D-C-485, PicoQuant) at an average power of 200 μW for hCypA and 100 μW for all other proteins. Single-molecule FRET efficiency histograms were acquired in samples with a protein concentration of about 20 to 50 pM, with the laser in either continuous-wave mode or pulsed mode at a repetition rate of 64 MHz; photon counts were recorded with a resolution of 16 ps by the counting electronics (time resolution was thus limited by the timing jitter of the detectors). For rapid mixing experiments (R17 at low concentrations of GdmCl), microfluidic mixers fabricated by replica molding in PDMS were used as described previously (4, 5). The measurements were performed in 50 mM sodium phosphate buffer, pH 7.0, 150mM β-mercaptoethanol (Sigma), 20mM cysteamine hydrochloride (Sigma), and 0.001% Tween 20 (Pierce) with varying concentrations of GdmCl (Pierce) for Csp*Tm*, R15, and R17. The measurements of *h*CypA were performed in 50 mM TrisHCl, 10 mM $MgCl_2$, 5 mM KCl, 100 mM β-mercaptoethanol and 0.001% Tween 20 (Pierce). For experiments in the microfluidic device, the Tween 20 concentration was increased to 0.01% to avoid surface adhesion of the proteins. All measurements were performed with instruments that were calibrated with Alexa Fluor 488 and Alexa Fluor 594 as described previously (6). Independent measurements of Cyp111 at two different instruments lead to an uncertainty of 0.02 in the mean transfer efficiency. Examples of single-molecule transfer efficiency histograms are shown in Fig. S1-3.

Two-focus fluorescence correlation spectroscopy (2fFCS).

2fFCS measurements (7) of donor-labeled hCypV2C were performed at 22 °C on a Micro Time 200 confocal microscope equipped with a differential interference contrast prism. The donor dye was excited alternatingly with two orthogonally polarized diode lasers at 483 nm (LDH-D-C-485, PicoQuant) with a repetition rate of 40 MHz and a laser power of 30 μW each. The concentration of labeled protein was 500 pM in 50 mM TrisHCl, 10 mM $MgCl_2$, 5 mM KCl, 100 mM β-mercaptoethanol, 0.001% Tween 20 (Pierce), pH 7.5 (native buffer) and varying concentrations of GdmCl. The distance between the two foci was determined using four standards, Oregon Green in water and 0.001% Tween20, and AlexaFluor488-labeled Csp*Tm*C67 (Csp-A488), *h*CypV2C (Cyp-A488), and monomeric GroEL-single ring (SR1-A488) in 5.07 M GdmCl, 50 mM sodium phosphate, 100 mM β-Mercaptoethanol, 0.001% Tween 20, pH 7.25. The reference value for the hydrodynamic radii (R_H) of Oregon Green is 0.6 nm (8). The reference values of the labeled proteins were determined under identical conditions using dynamic light scattering (DLS) with a Mambo-Laser 594nm (Cobolt, Sweden) at 100mW, resulting in 2.39 nm for Csp-A488, 3.71 nm for Cyp-A488, and 6.91 nm for SR1-A488. The focal distance was determined by iteratively minimizing the sum of the squared distances between reference R_H-value and the value determined by 2f-FCS. The fit converged to a focal distance of 442 nm, resulting in R_H-values for our reference substances of 0.47 nm (Oregon Green), 2.39 nm (Csp-A488), 3.6 nm (*h*Cyp-A488) and 6.98 nm (SR1-A488) (Fig. S4). Guanidinium chloride concentrations were measured with an Abbe refractometer (Krüss, Germany), and viscosities of the solutions were measured with a digital viscometer (DV-I+, Brookfield Engineering, Middleboro, MA, USA) with a CP40 spindle at 100 rpm.

Determination of R_G from mean transfer efficiencies.

In order to relate the distribution $P\left(r_G,\varepsilon,R_{G\Theta}\right)$ to a distance distribution $P\left(r,\varepsilon,R_{G\Theta}\right)$, which is required to describe the transfer efficiencies $\langle E\rangle$ of the polypeptide chains, we used as an approximation the conditional probability function $P\left(r|r_G\right)$ (9) that describes the distance distribution of two random points inside a sphere with the radius $\delta{\cdot}r_G$

$$P\left(r|r_G\right)=\frac{1}{\delta\cdot r_G}\left[3\left(\frac{r}{\delta\cdot r_G}\right)^2-\frac{9}{4}\left(\frac{r}{\delta\cdot r_G}\right)^3+\frac{3}{16}\left(\frac{r}{\delta\cdot r_G}\right)^5\right]\qquad 0\le r<2\delta\cdot r_G \tag{S1}$$

The actual value of δ is independent of the length of the polymer and was obtained from the condition that $6\left\langle R_G^2\right\rangle=\left\langle r^2\right\rangle$ at the Θ-state ($\delta=\sqrt{5}\approx 2.23$). Given Eqs. 1 and S1, the transfer efficiency between donor and acceptor results as

$$\langle E\rangle = \int_0^L E(r)\,P(r,\varepsilon,R_{G\Theta})\,dr = \int_0^L E(r) \int_{R_C}^{L/2} P(r|r_G)\,P(r_G,\varepsilon,R_{G\Theta})\,dr_G dr\,, \qquad \text{(S2)}$$

where $R_C = \left[3(N+1)\mathrm{v}/4\pi\right]^{1/3}$ is the radius of gyration of the most compact state, v is the weighted mean volume of one amino acid (v = 0.13nm^3) (10), and N is the number of peptide bonds between the fluorophores. Using two different guess values for $R_{G\Theta}$, we obtain two estimates for the root mean squared radius of gyration R_G, R_{G1} and R_{G2}, from the transfer efficiency $\langle E\rangle$. Although the shapes of $P(r_G,\varepsilon,R_{G\Theta})$ and $P(r,\varepsilon,R_{G\Theta})$ do depend on the choice of $R_{G\Theta}$, R_G is largely independent of the specific value of $R_{G\Theta}$ (Fig. S8). As guess values for the Θ-state, we assumed $R_{G\Theta,1} = \sqrt{l_p b/3}N^{1/2}$ with l_p = 0.4 nm as persistence length (Gaussian chain) (11) and $R_{G\Theta,2} = 0.658\mathrm{v}^{1/3}(N+1)^{1/2}$ (12). The volume fraction ϕ in Eq. 1 is given by $\phi = R_C^3 / R_G^3$. After calculating ε_i, with i = 1 for $R_{G\Theta,1}$ and i = 2 for $R_{G\Theta,2}$, the mean radii of gyration were obtained according to

$$R_{G,i} = \left(\int_{R_C}^{L/2} r_G^2 P(r_G,\varepsilon_i,R_{G\Theta,i})\,dR_G \right)^{1/2}. \qquad \text{(S3)}$$

The scaling exponents were determined from the segment length dependence of $R_G = (R_{G,1} + R_{G,2})/2$. The root mean squared difference σ_{12} between $R_{G,1}$ and $R_{G,2}$ was calculated as $\sigma_{12} = \sqrt{d^{-1}\sum_{j=1}^{d}\left(R_{G,1}(j) - R_{G,2}(j)\right)^2}$, where $R_{G,1}(j)$ and $R_{G,2}(j)$ are the radii of gyration at the GdmCl concentration j, and d is the total number of measurements. We found 0.05 nm ≤ σ_{12} ≤ 0.2 nm for all proteins and variants of this study, suggesting a sufficiently exact determination of R_G. The correct value for $R_{G\Theta}$ was finally estimated from the conditions at which ν = 1/2.

Simulations of a self-avoiding chain with excluded volume.

Equation S1 assumes that the spatial distribution of chain monomers of a polymer is spherically symmetric. However, several authors showed that self-avoiding chains in good solvent exhibit substantial asymmetry (13-17). We simulated an off-lattice self-avoiding chain by successively adding monomers with a volume of 0.13 nm^3 and a bond length of 0.38 nm until we obtained a chain of 50 monomers. In case a monomer interfered sterically with any other monomer, except its neighbor in sequence, the chain was deleted, and a new chain was started. It has been shown that this approach leads to an unbiased self-avoiding chain (16) comparable to the conventionally used Pivot-algorithm. We simulated 10,000 chains, and calculated $\left\langle R_G^2\right\rangle^{1/2} \equiv R_G$ and the mean transfer efficiency between the first and the last monomer. To quantify the asymmetry of the simulated chains, we calculated the asphericity

(Δ) according to Dima & Thirumalai (13) and found Δ = 0.45, indicating a significant deviation from spherical symmetry (Fig. S5). For the radius of gyration, we found R_G = 1.68 nm as an exact result. When we computed R_G from the mean transfer efficiency of the simulated chains using Eq. S1-3, we obtained a value of R_G = 1.76 nm, nearly independent of the choice of the radius of gyration of the Θ-state, which implies that we are overestimating R_G by about 5% under good solvent conditions. This result cannot serve as a proof that the functional form of Eq. S1 always leads to good estimates for R_G, especially not at the critical point, but we expect this deviation to be even smaller in poor solvent, since the asphericity is expected to be smaller for compact globules (13).

Comparison of mean-field theories for homopolymers and heteropolymers.

When treating a heteropolymer with a mean-field approach, it is natural to replace the conventional interaction parameter ε by a sum of the mean-field of the backbone (ε_{bb}) and an energy of the specific side-chains that is averaged over all monomers (ε_{sc}). Such an approach would lead the functional form of the free energy being almost unaltered compared to the homopolymer case as exemplified by a comparison between the homopolymer theory of Sanchez (12) and a statistical field-theory for heteropolymer collapse by Bryngelson and Wolynes (18). From Eq. 56 on p. 984 of ref. (12) we find for the free energy of the homopolymer in units of kT

$$F_{Homo} = -\frac{N}{2}\phi\varepsilon + N\left(\frac{1-\phi}{\phi}\right)\log\left(1-\phi\right) + F_{elast}\,. \qquad \text{S4}$$

In the same nomenclature, the free energy for the heteropolymer reads as

$$F_{Hetero} = -\frac{N}{2}z\phi\left(2\varepsilon + \Delta\varepsilon^2\right) + N\left(\frac{1-\phi}{\phi}\right)\log\left(1-\phi\right) + F_{elast} \qquad \text{S5}$$

with z being the coordination number, $\Delta\varepsilon$ being the variation of the mean-field interaction energy due to the heteropolymeric nature in the random energy approximation (REM), and F_{elast} is the elastic free energy resulting from the chain entropy (Eq. 23, p. 180 in ref. (18)). Both equations differ mainly in the interaction term.

Determination of scaling exponents.

In the power-law relation $R_G = \rho_0 N^{\nu}$ usually employed to describe the length scaling of polymers, ρ_0 cannot be assumed to be independent of solvent quality. An estimate for the dependence of ρ_0 on solvent quality can be obtained from chain statistics and the definition of R_G when following Flory (19) and Hammouda (20). The mean-squared distance between two monomers i and j for a freely joined chain with bond length b and persistence length l_p is

$$\left\langle r_{ij}^2 \right\rangle = 2l_p b\left|i-j\right|. \qquad \text{S6}$$

For a self-avoiding chain, Eq.S6 can be generalized to

$$\left\langle r_{ij}^2 \right\rangle = 2l_{p,ij}^* b\left|i-j\right|^{2\nu}. \qquad \text{S7}$$

Here $l_{p,ij}{}^*$ is a persistence length that depends on the solvent quality and the inter-dye distance between residues *i* and *j*. $l_{p,ij}{}^*$ also depends on the inter-dye distance because the tails for a given pair of residues *i* and *j* within the chain can alter the end-to-end distance. For the sake of simplicity, the persistence length is assumed to be independent of the specific positions *i* and *j* ($l_{p,ij}{}^* \approx l_p{}^*$), which is, strictly speaking, only true in ideal and good solvents. According to the definition of the radius of gyration,

$$R_G^2 = \frac{1}{2n^2}\sum_{i,j}\left\langle r_{ij}^2 \right\rangle, \qquad \text{S8}$$

with $n=N+1$ being the number of monomers in the chain. With Eq. S7, this yields

$$R_G^2 = \frac{2l_p^* b}{2n^2}\sum_{i,j}\left|i-j\right|^{2\nu} = \frac{2l_p^* b}{2n^2}\sum_{k=1}^{n} 2\left(n-k\right)k^{2\nu} = \frac{2l_p^* b}{n}\sum_{k=1}^{n}\left(1-\frac{k}{n}\right)k^{2\nu}. \qquad \text{S9}$$

Substituting $x = k/n$ and taking the limit of large *n*, the last expression can be written as

$$R_G^2 = 2l_p^* b n^{2\nu}\int_0^1 \left(1-x\right)x^{2\nu}dx = 2l_p^* b n^{2\nu}\left(\frac{1}{2\nu+1}-\frac{1}{2\nu+2}\right) \qquad \text{S10}$$

and we finally obtain for the radius of gyration of a self-avoiding chain

$$R_G = \sqrt{\frac{2l_p^* b}{(2\nu+1)(2\nu+2)}}\,n^{\nu}, \qquad \text{S11}$$

as given in ref. (20). A similar derivation for the freely joined chain can be found in Flory's book (19). Fitting the data of Kohn *et al.* (21) with Eq. S11 yields $\nu = 0.58$ and $l_p{}^* = 0.40 \pm 0.06$ nm (using $b = 0.38$ nm), in agreement with the value of 0.369 nm predicted from random sampling of the (ϕ,ψ) maps (22). A fit of the 10905 folded proteins from the pdb gives $\nu = 0.34$ and $l_p{}^* = 0.53$ nm. The origin of the higher value of $l_p{}^*$=0.53 nm in folded proteins compared to unfolded proteins in high denaturant might be a result of the specific secondary structure elements (α-helix, β-sheets) present in folded proteins or of the assumption that tail-effects are negligible, which is a very strong assumption for folded proteins. Analysis of our data with $l_p{}^*$=0.53 nm instead of $l_p{}^*$=0.40 nm results in critical exponents that are by a value of 0.04 lower than with $l_p{}^*$=0.40 nm. However, this does not affect our conclusions since the critical exponents for all proteins, except for cyclophilin, are still > 0.41. For cyclophilin, we obtain ν = 0.37 with $l_p{}^* = 0.53$ nm instead of ν = 0.40 with $l_p{}^* = 0.40$ nm. With Eq. S11,

l_p^*=0.40 nm and neglecting unity compared to N_{bonds}, the radius of gyration at the critical point is $R_{G\Theta} \approx 0.22\,\text{nm}\ N_{bonds}^{1/2}$.

Determination of the free energies of transfer, Δg_{sol}.

The δg_{sol} values (23) for the transfer of the individual amino acids from water to GdmCl were taken from Pace (24). No experimentally determined values for δg_{sol} are published for the amino acids Ser, Glu, Asp, Lys, and Arg. We thus followed the approach of O'Brien *et al.* (25) and approximated the values of Ser, Glu, and Asp by those of Thr, Gln, and Asn. The values of Lys and Arg were taken from O'Brien *et al.* (25). For interpolation, the δg_{sol} values were fitted with the Schellman weak binding model (26)

$$\delta g_{sol} = -\gamma\beta^{-1}\log\left(1+Ka\right). \tag{S12}$$

Here, γ is the number of bound GdmCl molecules, K is the binding constant, β is $(RT)^{-1}$, with R being the ideal gas constant and T being the temperature; a is the GdmCl activity (27). The δg_{sol} values, together with the values obtained for γ and K, are shown in Table S2. The fits with Eq. S12 are shown in Fig. S6. Finally, the average free energy of transfer per residue of an amino acid sequence from water to GdmCl is given by

$$\Delta g_{sol} = \delta g_{sol,b} + \sum_i p_i \delta g_{sol,i}\ , \tag{S13}$$

where $\delta g_{sol,i}$ is the free energy of transfer of an amino acid side chain of type i, p_i is the frequency of an amino acid of type i in the sequence, and $\delta g_{sol,b}$ is the free energy of transfer of one peptide bond. The summation is over all types of amino acids. We estimated the δg_{sol}-values for Asp and Glu, $\delta g_{sol}^{D,E}$ (Table S2), from the difference between the transfer free energy of ProTα in which all values of δg_{sol} for Glu and Asp residues were replaced by those for Gly, $\Delta g_{sol}^{D,E\to G}$, and the fit of $\Delta\varepsilon_{total}$ with Eq. S12, $\Delta\varepsilon_{total,Fit}$. Our estimate of $\delta g_{sol}^{D,E}$ is therefore given by

$$\delta g_{sol}^{D,E} = \frac{n_{total}}{n_{D,E}}\left(\Delta\varepsilon_{total,Fit} - \Delta g_{sol}^{D,E\to G}\right), \tag{S14}$$

with n_{total} = 129 being the total number of amino acids of ProTα and $n_{D,E}$ = 52 being the number of Asp and Glu in the sequence of ProTα (Fig. S7).

The effect of the fluorophore linkers on the scaling exponents.

The linker of the attached fluorophores might have an effect on the determined R_G-values and therefore also on the scaling exponents. We assumed l = 9 additional bonds for the linkers of our dyes, based on MD-simulations (28, 29) and previous work (11). However, since we have no information about the behavior of the linker and the dye at different denaturant

concentrations, we analyzed our data set for cyclophilin, which shows the most prominent collapse, with different values for the linker length l ranging from 3 to 18 bonds and found a variation of ν in water from 0.398 for the longest linker (l = 18) to 0.409 for the shortest linker (l = 3), which indicates a marginal effect of the linker length on the distance ranges mapped in our experiments (Fig. S9). In addition, we checked the effect of a fixed linker length that does not depend on solvent quality and analyzed the same data using

$$R_G = \left[\left(\frac{2 l_p^* b}{(2\nu+1)(2\nu+2)} \right)^{3/2} N^{3\nu} + R_{G,L}^3 \right]^{1/3} \qquad \text{S15}$$

with $R_{G,L}$ being an estimate for the linker length. Equation S15 results from the assumption that the volume of gyration of the protein-dye construct is the sum of the individual radii of gyration of chain and dye ($V_G = V_{G,Chain} + V_{G,Linker}$). Since the estimate for the additional distance introduced by the two dyes is approximately 1.47 nm (28), we estimated $R_{G,L}$ = 0.6 nm. To obtain an upper bound for the effect, we also used $R_{G,L}$ = 1.2 nm, which is twice the hydrodynamic radius of rhodamine, an analog of our fluorophores (8). We found the resulting effect of $R_{G,L}$ on R_G to be negligibly small (Fig. S9), again implying that the size of the dyes and their linkers do not affect the determined critical exponents.

Scaling of intra-chain energies with chain length.

By minimizing the free energy of the chain in the Sanchez model (Eq. 1 main text) and truncating the series expansion after the three-body interaction term, one obtains

$$\alpha^5 - \alpha^3 - \frac{c_1}{\alpha^3} = c_2 n^{1/2} \left(1 - \varepsilon\right), \qquad \text{(S16)}$$

where c_1 and c_2 are constants, and $n = N+1$ is the number of amino acids (12). Based on Eq. S16, the difference in the intra-chain interaction energy $\Delta\varepsilon = \varepsilon(n,a_{GdmCl,1}) - \varepsilon(n,a_{GdmCl,2})$ between two conditions with the GdmCl activities $a_{GdmCl,1}$ and $a_{GdmCl,2}$, corresponding to expansion factors α_1 and α_2, is given by

$$\Delta\varepsilon\left(n\right) = c_2^{-1} n^{-1/2} \left[\left(\alpha_2^5 - \alpha_2^3 - c_1 \alpha_2^{-3} \right) - \left(\alpha_1^5 - \alpha_1^3 - c_1 \alpha_1^{-3} \right) \right] = \Delta A n^{-1/2} . \qquad \text{(S17)}$$

The ratio of $\Delta\varepsilon(n_{DA})/\Delta\varepsilon_{total}(n_{total})$is

$$\frac{\Delta\varepsilon\left(n_{DA}\right)}{\Delta\varepsilon\left(n_{total}\right)} = \left(\frac{n_{total}}{n_{DA}} \right)^{1/2} \left(\frac{\Delta A_{DA}}{\Delta A_{total}} \right). \qquad \text{(S18)}$$

For the variant of a given protein with a sequence separation n_{DA} between the two fluorophores, the difference in α, $\Delta\alpha(n_{DA}) = \alpha_1(n_{DA}) - \alpha_2(n_{DA})$, between water, $\alpha_1(n_{DA})$, and a

GdmCl activity of 6, $\alpha_2(n_{DA})$, is very similar to the difference in α for the longest variant of the same protein $\Delta\alpha(n_{DA,longest})$. We obtained ratios $\Delta\alpha(n_{DA})/\Delta\alpha(n_{DA,longest})$ of 1.16 for *h*CypA, 1.03 for Csp*Tm,* 1.07 for R15 and R17, implying that $\Delta A_{DA}/\Delta A_{total} \approx 1$ for these proteins. For the IDPs prothymosin α and HIV integrase, $\Delta\alpha(n_{DA})/\Delta\alpha(n_{DA,longest})$ could not be calculated because data were only obtained for either one variant (HIV integrase) or two variants with almost identical sequence separation between the fluorophores (prothymosin α). Based on our results for the foldable proteins (*h*CypA, CspTm, R15 and R17), we assumed $\Delta A_{DA}/\Delta A_{total} \approx 1$ in these cases. We therefore obtain

$$\Delta\varepsilon_{total}(n_{total}) \approx \Delta\varepsilon(n)\left(\frac{n_{DA}}{n_{total}}\right)^{1/2} . \qquad \text{(S19)}$$

The remaining differences in $\Delta\varepsilon_{total}(n_{total})$ for the different variants of one protein in Fig. 5 might result from small deviations of $\Delta A_{DA}/\Delta A_{total}$ from one.

Link between unfolded-state collapse and folding.

To introduce a link between collapse and folding, we start from the probability distribution of chains with a given volume fraction ϕ as given by Sanchez Eq. 56, p. 984 (12)

$$P(\phi) = Z^{-1}\left(\frac{\phi_0}{\phi}\right)\exp\left\{-\frac{7}{2}\left(\frac{\phi}{\phi_0}\right)^{2/3} + n\left[\frac{\phi}{2}\varepsilon - \frac{1-\phi}{\phi}\ln(1-\phi)\right]\right\} \text{ with } \int_0^1 P(\phi)\,d\phi = 1 \qquad \text{(S20)}$$

Figure S10A shows several examples of $P(\phi)$ for different values of ε. We now assume that only unfolded proteins with a minimum volume fraction of $\phi > \phi_f$ can fold (Fig. S10A). One could imagine that the formation of a folding nucleus of critical size requires a minimum volume fraction ϕ_f. We further assume that chains with $\phi > \phi_f$ always fold completely to the native state, implying that the free energy of the folded state is always much smaller than that of the chains with $\phi > \phi_f$. The fraction of folding-competent collapsed chains, f_C, with $\phi > \phi_f$, and the fraction of expanded folding-incompetent chains, f_E, are then given by

$$f_C = \int_{\phi_f}^{1} P(\phi)\,d\phi \quad \text{and} \quad f_E = \int_{0}^{\phi_f} P(\phi)\,d\phi \qquad \text{(S21 a,b)}$$

(Fig. S10B) and the free energy difference between collapsed and expanded chains in units of k_BT is

$$\Delta F_{C-E} = -\ln\frac{f_C}{f_E} . \qquad \text{(S22)}$$

Figure S10C shows examples of $\Delta F_{C\text{-}E}$ for different sets of parameters and we find that $\Delta F_{C-E} \propto -\varepsilon$ for $\varepsilon < 1$ (Fig. S10C). Ziv & Haran (9) found a correlation between the $m_{N\text{-}U}$-value for the denaturant-induced unfolding of proteins (where $m_{N-U} = \partial \Delta F_{N-U} / \partial [D]$, and [$D$] is the concentration of denaturant) and the change in free energy of the unfolded chain with respect to a collapsed state, $m_{C-U} = \partial \Delta F_{C-U} / \partial [D]$. The quantity $\Delta F_{C\text{-}U}$ is identical to the quantity $\Delta F_{C\text{-}E}$. According to the result shown in Fig. 5A in the main text, we can substitute the intra-chain energy by the mean Tanford transfer values of the amino acid sequence,

$$\varepsilon = \varepsilon_0 - \gamma \ln\left(1 + Ka_{GdmCl}\right) \tag{S23}$$

and obtain with $\Delta F_{C-E} \propto -\varepsilon$

$$\Delta F_{C-E} \propto -\varepsilon_0 + \gamma \ln\left(1 + Ka_{GdmCl}\right). \tag{S24}$$

With the approximation that $\gamma \ln\left(1 + Ka_{GdmCl}\right) \approx m_T [D]$ (with $m_T > 0$), Eq. S24 leads to

$$\frac{\partial \Delta F_{C-E}}{\partial [D]} \propto m_T . \tag{S25}$$

The change in free energy difference between a collapsed ($\phi > \phi_f$) and an expanded state ($\phi < \phi_f$) is proportional to the change in free energy of transfer of the pure amino acids from water into a GdmCl-solution. When we use the typical Tanford expression for the free energy difference between folded and unfolded proteins ($\Delta F_{N\text{-}U}$), as for example given in Eq. 2 by Ziv & Haran (9), and set $\Delta F_{N\text{-}U} = \Delta F_{N\text{-}E}$, we have

$$\Delta F_{N-E}(D) = \Delta F_{N-E}(0) + n m_T [D] \Delta\alpha \quad \text{and} \quad \frac{\partial \Delta F_{N-E}}{\partial [D]} = n m_T \Delta\alpha \tag{S26 a,b}$$

with $\Delta\alpha = \alpha_E - \alpha_N$ being the average difference in solvent accessible surface area between the expanded unfolded and the folded state. Since $\Delta\alpha$ is a constant, it is clear by comparing Eq. S26b with Eq. S25 that

$$\frac{\partial \Delta F_{C-E}}{\partial [D]} \propto \frac{\partial \Delta F_{N-E}}{\partial [D]}, \tag{S27}$$

which is the correlation found by Ziv & Haran (9).

Interpolation and Extrapolation of the experimentally determined R_G-values.

To obtain R_G-values for the different inter-dye variants of our proteins at identical concentrations of GdmCl, all raw-data sets (R_G vs. GdmCl-concentration) were fitted with the empirical equation

$$R_G = R_{G0} + \frac{a_1[GdmCl]}{K + [GdmCl]} + a_2 \exp\left(-a_3[GdmCl]\right), \tag{S.28}$$

where the third term describes the re-expansion of the IDP's integrase and prothymosin at very low concentrations of GdmCl. For all foldable proteins $a_2 = 0$. The fits of the raw data are shown in Fig. S11. The values of the fits with Eq. S28 were used to obtain the results shown in Fig. 2 in the main text. The data below 0.6 M GdmCl ($a_{GdmCl} < 0.19$) for all *Csp*-variants, below 0.2 M GdmCl ($a_{GdmCl} < 0.033$) for all *Cyp*-variants, and below 0.3 M ($a_{GdmCl} < 0.07$) for $R_{15}60$ and $R_{17}93$ were extrapolated to 0 M GdmCl using Eq. S28. For $R_{17}60$, the unfolded state was also investigated in a micro-fluidic device (5) down to 0.03 M GdmCl.

Calculation of scaling exponents from net charge and hydrophobicity.

The correlation between scaling exponent and net charge Q and the mean hydrophobicity H (Fig. 6*A, B* main text) where fit with the empirical equations

$$\nu(Q) = 1/3 + a\left[1 + \exp\left(x_0 - Q\right)/z\right]^{-1} \text{ and } \nu(H) = 1/3 + a\left[1 + \exp\left(x_0 + cH - d\right)/z\right]^{-1} \tag{S29}$$

where we assumed a negative correlation between the mean net charge Q and the mean hydrophobicity H according to $Q = -cH + d$. The equation provides reasonable limits for ν,

$$\lim_{H\to 1} \nu(H) = 1/3$$

$$\lim_{H\to 0} \nu(H) = 0.71$$

$$\lim_{Q\to 1} \nu(Q) = 0.71.$$

The parameters obtained are $a = 0.394$, $z = 0.09$, $x_0 = 0.114$, $c = 1.72$, and $d = 0.9$. In order to combine the two different correlations of ν with net charge, $\nu(Q)$, and ν with hydrophobicity, $\nu(H)$, (Fig. 6*A, B*, main text), we used polyampholyte theory (3, 30) to decide which correlation is most suited to predict the scaling exponent of a given amino acid sequence. Polyampholyte theory provides an expression for the effect of charges on the excluded volume υ, expressed as an excess volume υ^*:

$$\upsilon^* = \frac{4\pi l_B \left(f-g\right)^2}{\kappa^2} - \frac{\pi l_B^2 \left(f+g\right)^2}{\kappa} \tag{S30}$$

with f being the fraction of positive charges in a chain with length n ($f = n_+/n$), g being the fraction of negative charges ($g = n_-/n$), $\kappa^{-1} = 0.304\,\text{nm}/\sqrt{I}$ being the Debye length at ionic strength I, and $l_B = e^2/\left(4\pi\varepsilon_0\varepsilon_r k_B T\right)$ being the Bjerrum length, where e is the elementary charge, ε_0 is the dielectric constant, ε_r is the permittivity of water, k_B is the Boltzmann constant, and T is the temperature. Values of υ^* greater than zero indicate a net electrostatic repulsion, in which case we use $\nu(Q)$ to estimate the scaling exponent, whereas $\upsilon^* \le 0$ indicates a net attraction, in which case we use $\nu(H)$ to estimate the scaling exponent. For I = 0.15 M and T = 298 K, we calculated υ^* for every sequence that was drawn randomly from the amino acid frequency distribution of ancestral proteins, current proteins, and proteins in distant time given by Table 3 in ref. (31). Whether $\nu(Q)$ or $\nu(H)$ should be used to estimate the scaling exponent ν was decided according to the following criterion:

$$\nu = \begin{cases} \nu(Q) & \upsilon^* > 0 \vee f = 0 \vee g = 0 \\ \nu(H) & \upsilon^* \le 0 \vee f = 0 \wedge g = 0 \end{cases}. \tag{S31}$$

Table S1. Proteins and variants used in this study

protein	variant	N[b]	mutation[a]	sequence
Csp*Tm*	Csp33	33	M34G/p.E33_E35*ins*RC/E69C	*GPG MRGKVKFFDS KKGYGFITKD EGGDVFVHFS AIEGR CEGF KTLKEGQVVE FEIQEGKKGG QAAHVKVVEC*
	ΔCsp33	33	M34G/p.G34_E35*ins*RC/E69C/p.M1_R35*del*	*CEGF KTLKEGQVVE FEIQEGKKGG QAAHVKVVEC*
	Csp46	46	E21C/E67C	*GPG MRGKVKFFDS KKGYGFITKD CGGDVFVHFS AIEMEGFKTL KEGQVVEFEI QEGKKGGQAA HVKVVEC*
	Csp57	57	S10C/E67C	*GPG MRGKVKFFDC K KGYGFITKDE GGDVFVHFSA IEMEGFKTL KEGQVVEFEI QEGKKGGQAA HVKVVEC*
	Csp66	66	p.M1_R2*ins*C/E68C	*GPG MCRGKVKFFD SKKGYGFITK DEGGDVFVHF SAIEMEGFKT LKEGQVVEFE IQEGKKGGQA AHVKVVEC*
R15	R_{15}60	60	A39C/S99C	*KLKEANKQQN FNTGIKDFDF WLSEVEALLA SEDYGKDLCS VNNLLKKHQL LEADISAHED RLKDLNSQAD SLMTSSAFDT SQVKDKRETI NGRFQRIKCM AAARRAKLNES HRL*
	R_{15}93	93	N6C/S99C	*KLKEACKQQN FNTGIKDFDF WLSEVEALLA SEDYGKDLAS VNNLLKKHQL LEADISAHED RLKDLNSQAD SLMTSSAFDT SQVKDKRETI NGRFQRIKCM AAARRAKLNES HRL*
R17	R_{17}60	60	A39C/K99C	*RLEESLEYQQ FVANVEEEEA WINEKMTLVA SEDYGDTLCA IQGLLKKHEA FETDFTVHKD RVNDVAANGE DLIKKNNHHV ENITAKMKGL KGKVSDLECA AAQRKAKLDE NSAFLQ*
	R_{17}93	93	L6C/K99C	*RLEESCEYQQ FVANVEEEEA WINEKMTLVA SEDYGDTLAA IQGLLKKHEA FETDFTVHKD RVNDVAANGE DLIKKNNHHV ENITAKMKGL KGKVSDLECA*

protein	variant	N[b]	mutation[a]	sequence
				AAQRKAKLDE NSAFLQ
***h*Cyp**	Cyp96	96	K28C/G124C	*GP MVNPTVFFDI AVDGEPLGRV SFELFADCVP KTAENFRALS TGEKGFGYKG SSFHRIIPGF MSQGGDFTRH NGTGGKSIYG EKFEDENFIL KHTGPGILSM ANAGPNTNGS QFFISTAKTE FLDCKHVVFG KVKEGMNIVE AMERFGSRNG KTSKKITIAD SGQLE*
	Cyp111	111	D13C/G124C	*GP MVNPTVFFDI AVCGEPLGRV SFELFADKVP KTAENFRALS TGEKGFGYKG SSFHRIIPGF MSQGGDFTRH NGTGGKSIYG EKFEDENFIL KHTGPGILSM ANAGPNTNGS QFFISTAKTE FLDCKHVVFG KVKEGMNIVE AMERFGSRNG KTSKKITIAD SGQLE*
	Cyp122	122	V2C/G124C	*GP MCNPTVFFDI AVDGEPLGRV SFELFADKVP KTAENFRALS TGEKGFGYKG SSFHRIIPGF MSQGGDFTRH NGTGGKSIYG EKFEDENFIL KHTGPGILSM ANAGPNTNGS QFFISTAKTE FLDCKHVVFG KVKEGMNIVE AMERFGSRNG KTSKKITIAD SGQLE*
	Cyp152	152	V2C/K154C	*GP MCNPTVFFDI AVDGEPLGRV SFELFADKVP KTAENFRALS TGEKGFGYKG SSFHRIIPGF MSQGGDFTRH NGTGGKSIYG EKFEDENFIL KHTGPGILSM ANAGPNTNGS QFFISTAKTE FLDGKHVVFG KVKEGMNIVE AMERFGSRNG KTSCKITIAD SGQLE*
	Cyp163	163	V2C/E165C	*GP MCNPTVFFDI AVDGEPLGRV SFELFADKVP KTAENFRALS TGEKGFGYKG SSFHRIIPGF MSQGGDFTRH NGTGGKSIYG EKFEDENFIL KHTGPGILSM ANAGPNTNGS QFFISTAKTE FLDGKHVVFG KVKEGMNIVE AMERFGSRNG KTSKKITIAD SGQLC*
IN	IN	56		*GSHC FLDGIDKAQE EHEKYHSNWR AMASDFNLPP VVAKEIVASC DKCQLKGEAM HGQVDC*

protein	variant	N[b]	mutation[a]	sequence
ProTα	ProTC2	53	S2C	*MAHHHHHHS AALEVLFQGP MCDAAVDTSS EITTKDLKEK KEVVEEAENG RDAPANGNAN EENGEQEADN EVDEECEEGG EEEEEEEEGD GEEEDGDEDE EAESATGKRA AEDDEDDDVD TKKQKTDEDD*
	ProTC110	54	D110C	*MAHHHHHHS AALEVLFQGP MSDAAVDTSS EITTKDLKEK KEVVEEAENG RDAPANGNAN EENGEQEADN EVDEECEEGG EEEEEEEEGD GEEEDGDEDE EAESATGKRA AEDDEDDDVD TKKQKTDEDC*

[a] Additional mutations Csp*Tm*: W7F/W29F; R17: C68A; *h*Cyp: W121F/C52S/C62S/C115S/C161S

[b] Number of peptide bonds between donor and acceptor attachment sites

Table S2. Free energies of transfer (δg_{sol}) and fit parameters for the single amino acis.

	$-\delta g_{sol}$ (cal mol^{-1})[b]					
	GdmCl (M)					
residue	**1**	**2**	**4**	**6**	γ	K[a]
Ala	10	20	30	45	0.030 ± 0.004	3 ± 1
Val	85	115	195	265	0.150 ± 0.026	5 ± 2
Leu	150	210	355	480	0.275 ± 0.042	5 ± 2
Ile	135	190	320	430	0.244 ± 0.036	5 ± 2
Met	150	245	400	535	0.317 ± 0.024	4 ± 1
Cys	150	245	400	535	0.317 ± 0.024	4 ± 1
Phe	215	355	580	775	0.462 ± 0.032	4 ± 1
Tyr	235	385	605	770	0.416 ± 0.018	6 ± 1
Trp	400	630	980	1235	0.640 ± 0.034	7 ± 1
Pro	100	140	240	320	0.184 ± 0.027	5 ± 2
Thr	65	90	120	125	0.042 ± 0.006	67 ± 48
His	180	285	385	420	0.167 ± 0.021	27 ± 13
Asn	200	320	490	645	0.344 ± 0.022	6 ± 1
Gln	135	215	315	360	0.163 ± 0.014	14 ± 4
Gly	0	0	0	0	0	0
backbone	83	134	207	245	0.121 ± 0.009	9 ± 2
Ser[c]	65	90	120	125	0.042 ± 0.006	67 ± 48
Asp[c]	200	320	490	645	0.344 ± 0.022	6 ± 1
Glu[c]	135	215	315	360	0.163 ± 0.014	14 ± 4
Lys[c]	68	136	272	408	0.394 ± 0.027	1.1 ± 0.2
Arg[c]	42	85	170	254	0.245 ± 0.017	1.1 ± 0.2
Glu, Asp[d]	-	112	439	798	3 ± 3	0.12 ± 0.15

[a] Values on GdmCl-activity scale; [b] from Pace(24); [c] estimates for δg_{sol} according to O'Brien *et al.*(25),[d] Values estimated in this study

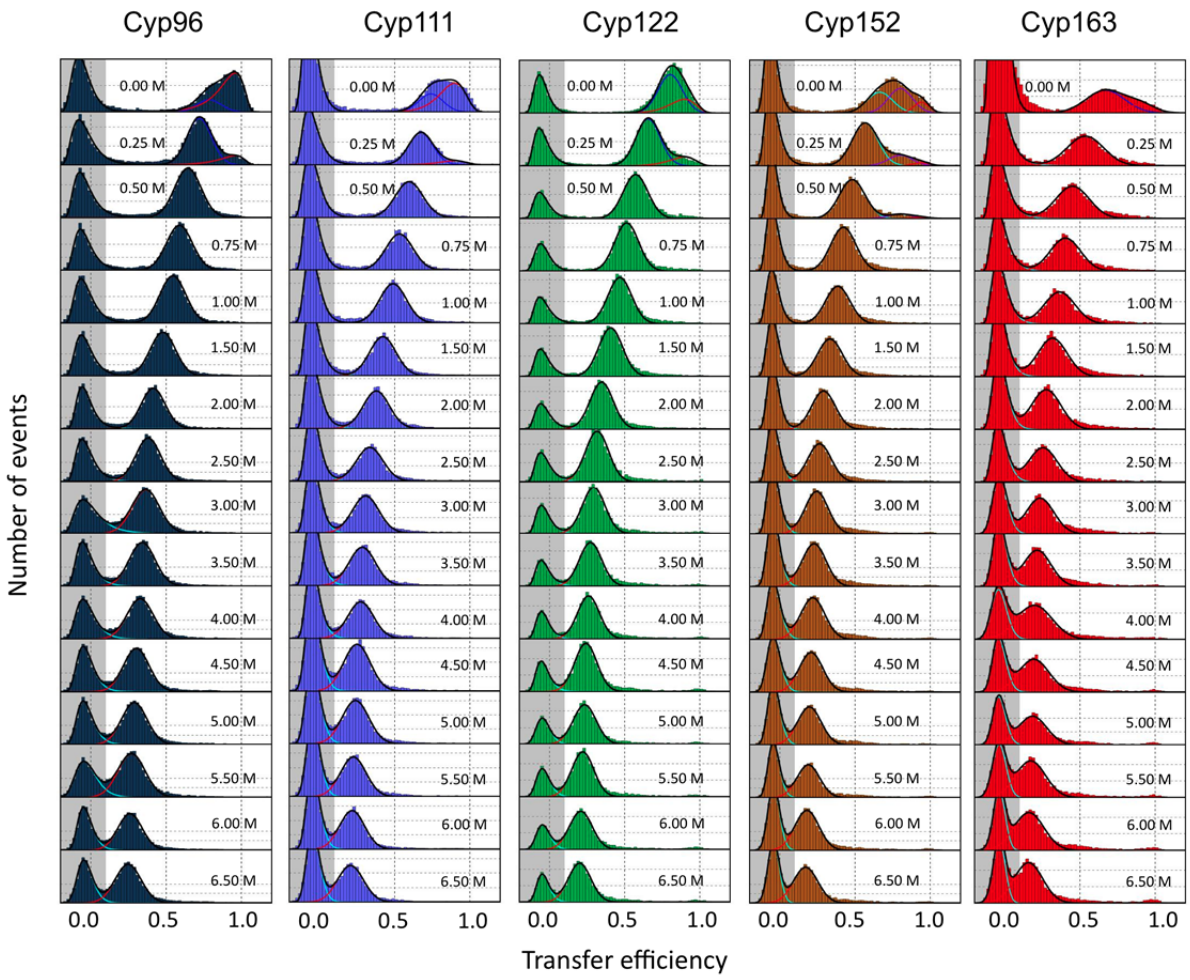

Figure S1. Transfer efficiency histograms of *h*Cyp variants at different concentrations of GdmCl. Solid lines are fits according to a sum of a Gaussian distribution describing the unfolded state population and two log-normal functions describing the native transfer efficiency distribution at high transfer efficiencies, and the donor-only population at low transfer efficiencies, respectively.

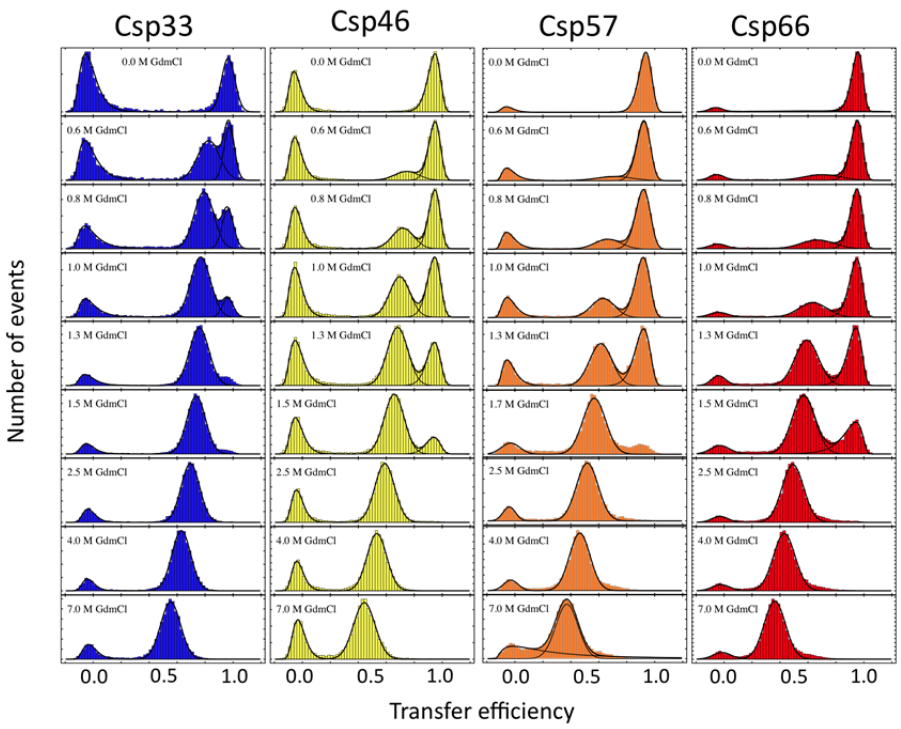

Figure S2. Selected transfer efficiency histograms of Csp*Tm* variants at different concentrations of GdmCl. Solid lines are fits according to a sum of a Gaussian distribution describing the unfolded state population and two log-normal functions describing the native transfer efficiency distribution at high transfer efficiencies, and the donor-only population at low transfer efficiencies, respectively.

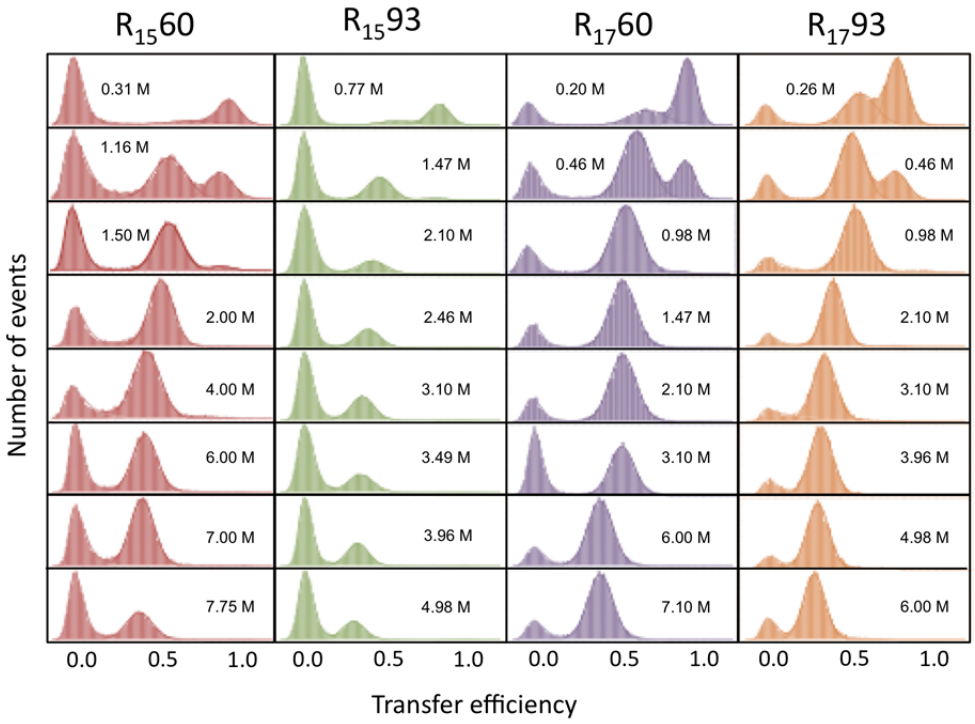

Figure S3. Selected transfer efficiency histograms of R15 and R17 variants at different concentrations of GdmCl. Solid lines are fits according to a sum of a Gaussian distribution describing the unfolded state population and two log-normal functions describing the native transfer efficiency distribution at high transfer efficiencies, and the donor-only population at low transfer efficiencies, respectively.

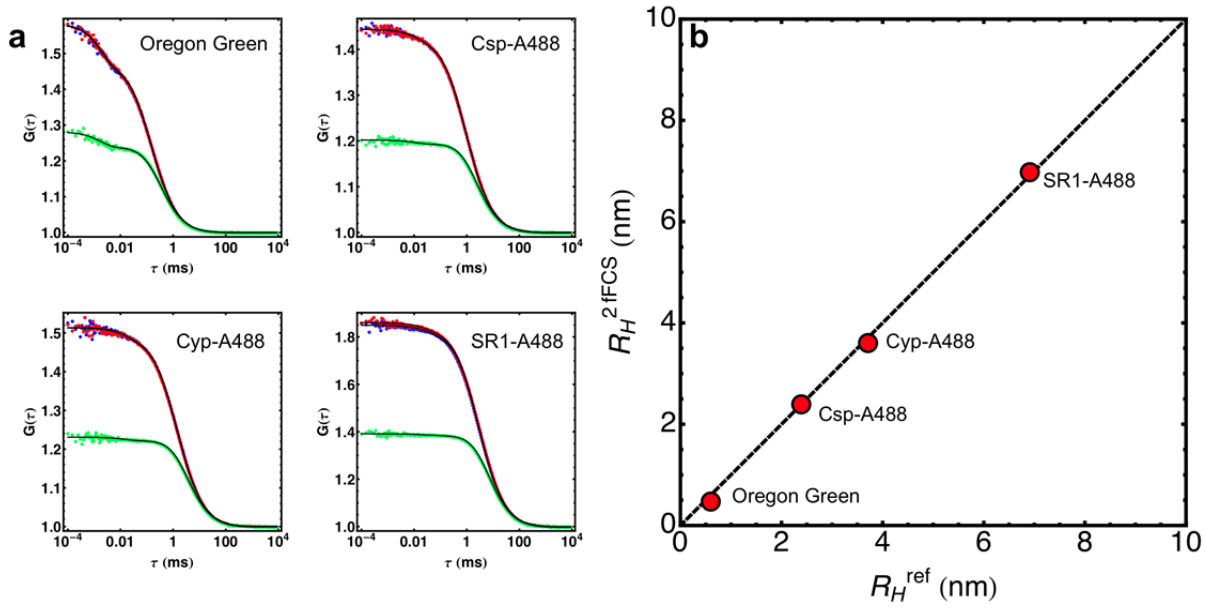

Figure S4. Calibration of 2f-FCS. (**a**) 2f-FCS autocorrelation functions (blue, red) and crosscorrelation functions (green) for Oregon Green in water and Csp-A488, Cyp-A488 and SR1-A488 in 5.07 M GdmCl. Solid black lines are fits according to Dertinger *et al.*(7). The fits include a component describing the triplett-lifetime of the fluorophores. The measurements were performed at 21.8 °C with a laser power of 30 µW for each focus. We obtained the following diffusion coefficients: 4.68×10^{-6} cm^2s^{-1} (η = 0.98 mPa s) Oregon Green, 6.54×10^{-7} cm^2s^{-1} (η = 1.38 mPa s) Csp-A488, 4.35×10^{-7} cm^2s^{-1} (η = 1.38 mPa s) Cyp-A488, 2.24×10^{-7} cm^2s^{-1} (η = 1.38 mPa s) SR1-A488 (**b**) Correlation between hydrodynamic radius measured with 2fFCS (R_H^{2fFCS}) and hydrodynamic radius reported in literature (R_H^{ref}) for Oregon Green and determined with DLS for Csp-A488, Cyp-A488 and SR1-A488 at 5.07 M GdmCl with a focal distance of 442 nm.

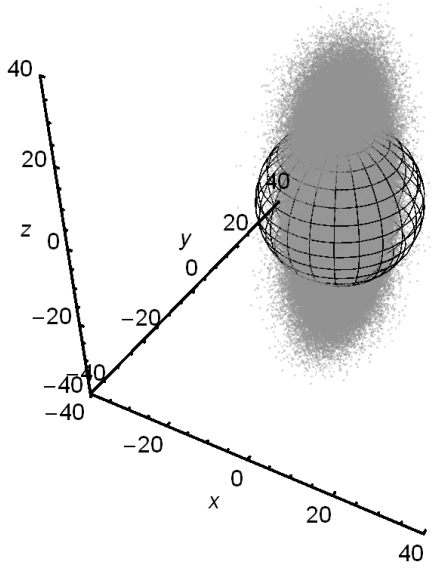

Figure S5. Graphical representation of the monomer coordinates of 2000 self-avoiding chains with R_G = 1.68 nm (gray) aligned along their principal axis. Each chain consists of 50 monomers. The sphere represents the model used in Eq. S1 for the determination of R_G from the mean transfer efficiency ($R_{G,FRET}$). The radius of the sphere is $R_{G,FRET}$ = 1.76 nm. The axis units are in Å.

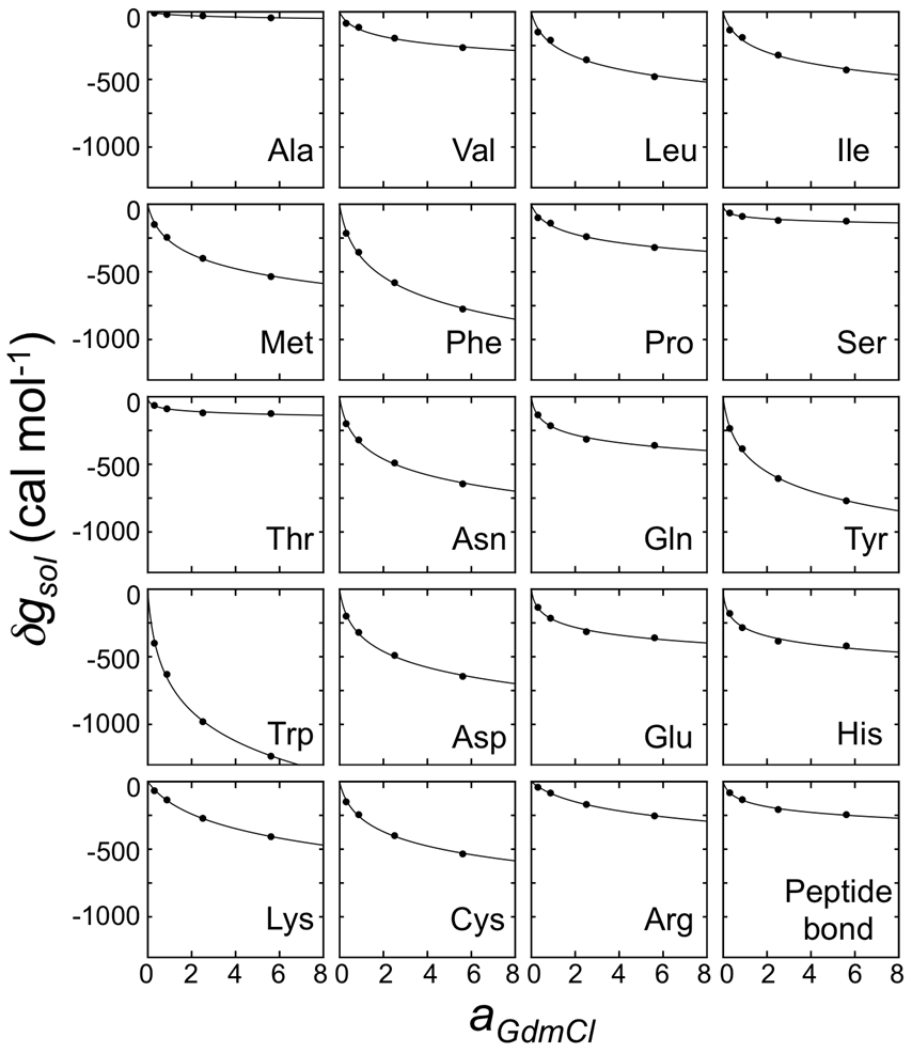

Figure S6. Fits of the free energies of transfer for the single amino acids δg_{sol} with the Schellman binding model (Eq. S12). The values for Glu and Asp are identical to that of Gln and Asn.

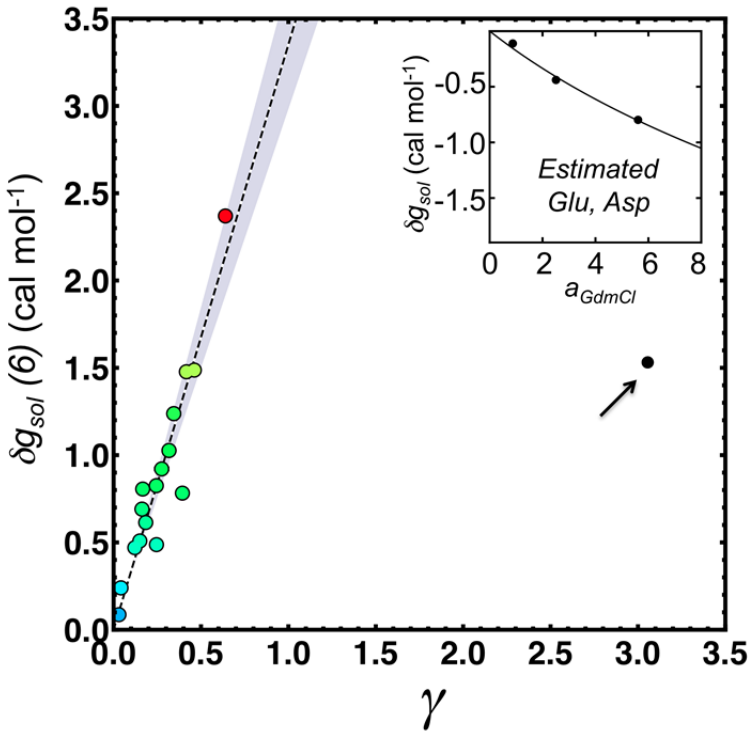

Figure S7. Correlation of the free energies of transfer for the single amino acids δg_{sol} at an GdmCl-activity of 6 with the number of GdmCl-binding sites γ. The black point (indicated by the arrow) is the value for Glu and Asp determined from the change in the intra-chain interaction free energy of ProT. The color scale increases from blue to red with increasing δg_{sol} . The red point results from Trp. Inset: Estimated change in the free energy of transfer for Glu and Asp. Parameters are given in Table S1.

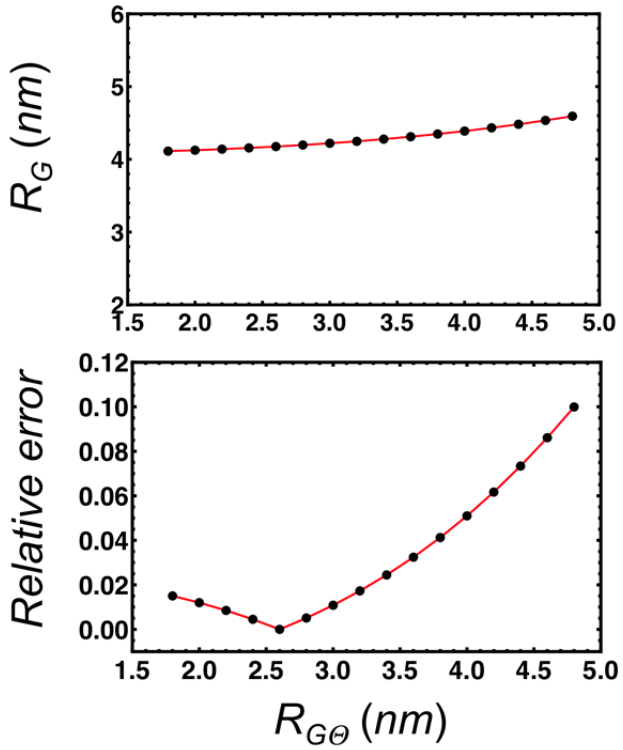

Figure S8. Change in R_G on varying guess values of $R_{G\Theta}$. Absolute R_G-values for Cyp163 at 6.3 M GdmCl as function of $R_{G\Theta}$ calculated using Eq. S2 (top). Relative error in estimating R_G as function of $R_{G\Theta}$ (bottom).

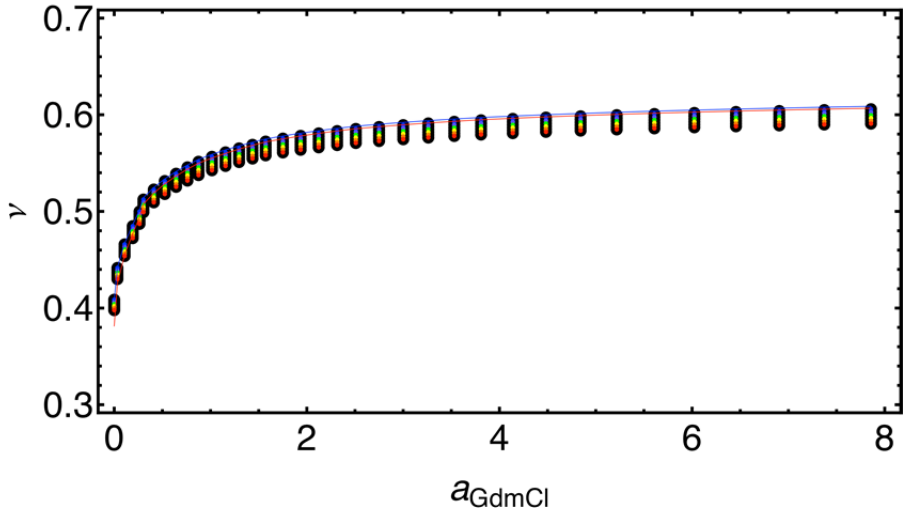

Figure S9. Critical exponents obtained for varying linker length (circles) with linker lengths corresponding to 3 (blue), 6 (lighter blue), 9 (green), 12 (yellow), 15 (orange) and 18 (red) equivalent bond length. The nearly indistinguishable red and blue lines correspond to an analysis with a fixed distance offset as given by Eq. S15.

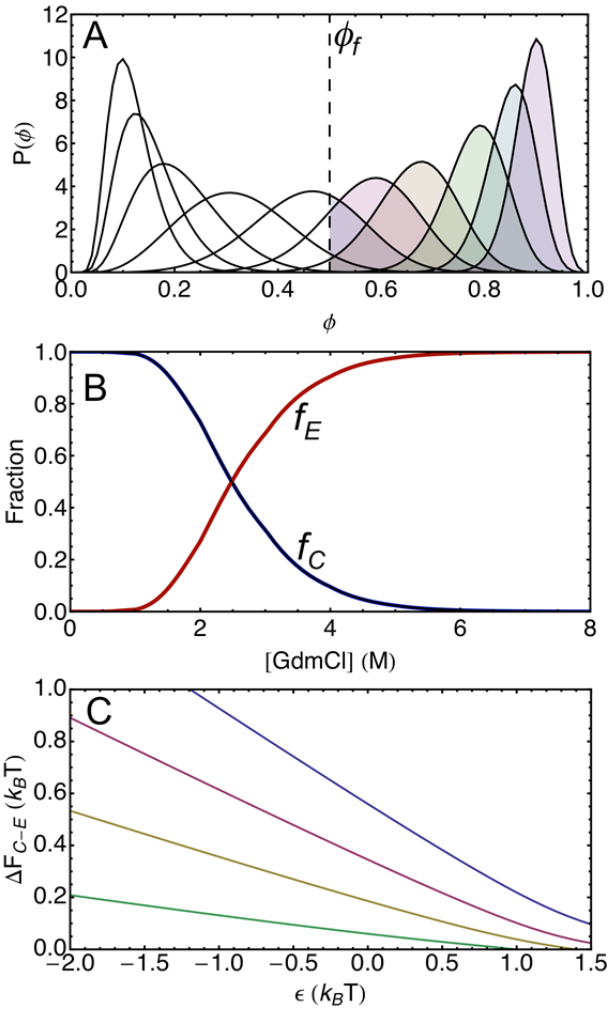

Figure S10. Volume fraction distributions $P(\phi)$ (Eq. S20) (A) and the fraction of collapsed (folding competent) and expanded (folding incompetent) chains as a function of the GdmCl concentration (Eq. S21 a,b) (B) and free energy difference between expanded and collapsed chains (C). (A) Colored areas indicate the fraction of chains with $\phi > \phi_f$ for chains with increasing intra-chain interaction energies (ε). (B) The parameter set was $\phi_0 = \phi_f = 0.29$, $n = 150$, $\varepsilon_0 = 2$, $\gamma = 0.3$, $K = 10$. (C) Calculated according to Eq. S22 with $n = 100$ and $\phi_0 = 0.29$ for different values of $\phi_f = 0.8$ (blue), $\phi_f = 0.6$ (red), $\phi_f = 0.4$ (yellow), $\phi_f = 0.2$ (green).

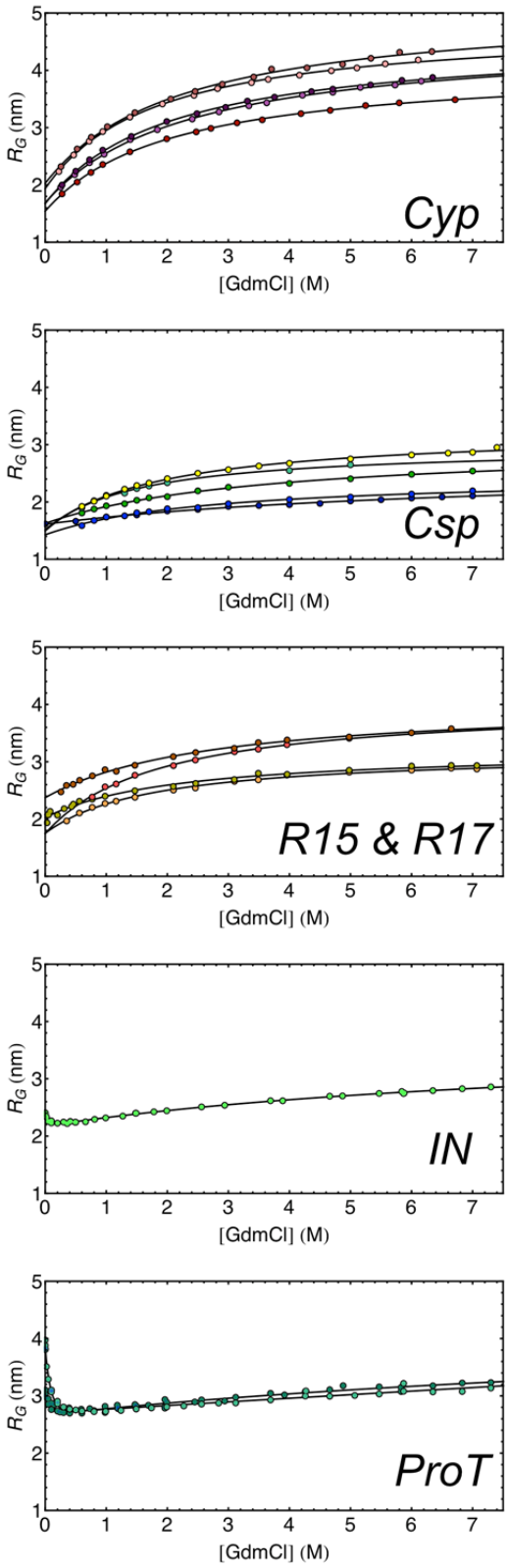

Figure S11. R_G-values determined from the mean transfer efficiencies using Eq. S1-3 and fits according to Eq. S28. The color code for the different variants is shown in Fig. 3B in the main text.

References

1. Scott K, Batey S, Hooton K, & Clarke J (2004) The Folding of Spectrin Domains I: Wild-type Domains Have the Same Stability but very Different Kinetic Properties. *J Mol Biol* 344(1):195-205.
2. Soranno A*, et al.* (2012) Quantifying internal friction in unfolded and intrinsically disordered proteins with single molecule spectroscopy. *Proc Natl Acad Sci USA* In Press.
3. Müller-Späth S*, et al.* (2010) Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc Natl Acad Sci USA* 107(33):14609-14614.
4. Hofmann H*, et al.* (2010) Single-molecule spectroscopy of protein folding in a chaperonin cage. *Proc Natl Acad Sci USA* 107(26):11793-11798.
5. Pfeil S, Wickersham C, Hoffmann A, & Lipman E (2009) A microfluidic mixing system for single-molecule measurements. *Rev Sci Instrum* 80(5):055105.
6. Schuler B (2007) Application of single molecule Förster resonance energy transfer to protein folding. *Methods Mol Biol* 350:115-138.
7. Dertinger T*, et al.* (2007) Two-focus fluorescence correlation spectroscopy: a new tool for accurate and absolute diffusion measurements. *Chemphyschem* 8(3):433-443.
8. Mueller CB*, et al.* (2008) Precise measurement of diffusion by multi-color dual-focus fluorescence correlation spectroscopy. *EPL* 83(4):46001-p46001-46005.
9. Ziv G & Haran G (2009) Protein Folding, Protein Collapse, and Tanford's Transfer Model: Lessons from Single-Molecule FRET. *J Am Chem Soc* 131(8):2942-2947.
10. Zamyatnin A (1984) Amino acid, peptide, and protein volume in solution. *Annu Rev Biophys Bioeng* 13:145-165.
11. Hoffmann A*, et al.* (2007) Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proc Natl Acad Sci USA* 104(1):105-110.
12. Sanchez I (1979) Phase Transition Behavior of the Isolated Polymer Chain. *Macromolecules* 12:980-988.
13. Dima R & Thirumalai D (2004) Asymmetry in the shapes of folded and denatured states of proteins. *J Phys Chem B* 108(21):6564-6570.
14. Theodorou DN & Suter UW (1985) Shape of Unperturbed Linear-Polymers - Polypropylene. *Macromolecules* 18(6):1206-1214.
15. Tran HT & Pappu RV (2006) Toward an accurate theoretical framework for describing ensembles for proteins under strongly denaturing conditions. *Biophys J* 91(5):1868-1886.
16. Hadizadeh S, Linhananta A, & Plotkin SS (2011) Improved Measures for the Shape of a Disordered Polymer To Test a Mean-Field Theory of Collapse. *Macromolecules* 44(15):6182-6197.
17. Sfatos CD, Gutin AM, & Shakhnovich EI (1994) Phase transitions in a "many-letter" random heteropolymer. *Phys Rev E* 50(4):2898-2905.
18. Bryngelson J & Wolynes P (1990) A Simple Statistical Field-Theory of Heteropolymer Collapse with Application to Protein Folding. *Biopolymers* 30:177-188.
19. Flory P (1989) *Statistical Mechanics of Chain Molecules* (Carl Hanser Verlag, Munich Vienna New York).

20. Hammouda B (1993) SANS from Homogeneous Polymer Mixtures - A unified Overview. *Adv Polym Sci* 106:87-133.
21. Kohn J*, et al.* (2004) Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc Natl Acad Sci USA* 101(34):12491-12496.
22. Zhou H (2002) Dimensions of denatured protein chains from hydrodynamic data. *J Phys Chem B* 106:5769-5775.
23. Nozaki Y & Tanford C (1970) The solubility of amino acids, diglycine, and triglycine in aqueous guanidine hydrochloride solutions. *J Biol Chem* 245(7):1648-1652.
24. Pace C (1986) Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol* 131:266-280.
25. O'Brien E, Ziv G, Haran G, Brooks B, & Thirumalai D (2008) Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model. *Proc Natl Acad Sci USA* 105(36):13403-13408.
26. Schellman J (2002) Fifty years of solvent denaturation. *Biophys Chem* 96(2-3):91-101.
27. Makhatadze G, Fernandez J, Freire E, Lilley T, & Privalov P (1993) Thermodynamics of Aqueous Guanidinium Hydrochloride Solutions in the Temperature-Range From 283.15 to 313.15-K. *J Chem Eng Data* 38(1):83-87.
28. McCarney ER*, et al.* (2005) Site-specific dimensions across a highly denatured protein; a single molecule study. *J Mol Biol* 352(3):672-682.
29. Schröder GF, Alexiev U, & Grubmüller H (2005) Simulation of fluorescence anisotropy experiments: probing protein dynamics. *Biophys J* 89(6):3757-3770.
30. Higgs P & Joanny J-F (1991) Theory of polyampholyte solutions. *J Chem Phys* 94(2):1543-1554.
31. Jordan IK*, et al.* (2005) A universal trend of amino acid gain and loss in protein evolution. *Nature* 433(7026):633-638.