How accurate are polymer models in the analysis of Förster resonance energy transfer experiments on proteins?

Edward P. O'Brien,[1,2] Greg Morrison,[1,3] Bernard R. Brooks,[2] and D. Thirumalai[1,4,a)]
[1]*Biophysics Program, University of Maryland, College Park, Maryland 20742, USA*
[2]*Laboratory of Computational Biology, National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, USA*
[3]*School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, 02138, USA*
[4]*Department of Chemistry and Biochemistry and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, USA*

(Received 26 September 2008; accepted 6 January 2009; published online 24 March 2009)

Single molecule Förster resonance energy transfer (FRET) experiments are used to infer the properties of the denatured state ensemble (DSE) of proteins. From the measured average FRET efficiency, $\langle E\rangle$, the distance distribution $P(R)$ is inferred by assuming that the DSE can be described as a polymer. The single parameter in the appropriate polymer model (Gaussian chain, wormlike chain, or self-avoiding walk) for $P(R)$ is determined by equating the calculated and measured $\langle E\rangle$. In order to assess the accuracy of this "standard procedure," we consider the generalized Rouse model (GRM), whose properties [$\langle E\rangle$ and $P(R)$] can be analytically computed, and the Molecular Transfer Model for protein L for which accurate simulations can be carried out as a function of guanadinium hydrochloride (GdmCl) concentration. Using the precisely computed $\langle E\rangle$ for the GRM and protein L, we infer $P(R)$ using the standard procedure. We find that the mean end-to-end distance can be accurately inferred (less than 10% relative error) using $\langle E\rangle$ and polymer models for $P(R)$. However, the value extracted for the radius of gyration (R_g) and the persistence length (l_p) are less accurate. For protein L, the errors in the inferred properties increase as the GdmCl concentration increases for all polymer models. The relative error in the inferred R_g and l_p, with respect to the exact values, can be as large as 25% at the highest GdmCl concentration. We propose a self-consistency test, requiring measurements of $\langle E\rangle$ by attaching dyes to different residues in the protein, to assess the validity of describing DSE using the Gaussian model. Application of the self-consistency test to the GRM shows that even for this simple model, which exhibits an order $\rightarrow$disorder transition, the Gaussian $P(R)$ is inadequate. Analysis of experimental data of FRET efficiencies with dyes at several locations for the cold shock protein, and simulations results for protein L, for which accurate FRET efficiencies between various locations were computed, shows that at high GdmCl concentrations there are significant deviations in the DSE $P(R)$ from the Gaussian model. © *2009 American Institute of Physics*. [DOI: 10.1063/1.3082151]

I. INTRODUCTION

Much of our understanding of how proteins fold comes from experiments in which folding is initiated from an ensemble of initially unfolded molecules whose structures are hard to characterize.[1] In many experiments, the initial structures of the denatured state ensemble (DSE) are prepared by adding an excess amount of denaturants or by raising the temperature above the melting temperature (T_m) of the protein.[2] Theoretical studies have shown that folding mechanisms depend on the initial conditions, i.e., the nature of the DSE.[3] Thus, a quantitative description of protein folding mechanisms requires a molecular characterization of the DSE—a task that is made difficult by the structural diversity of the ensemble of unfolded states.[4,5]

In an attempt to probe the role of initial conditions on folding, single molecule Förster resonance energy transfer (FRET) experiments are being used to infer the properties of unfolded proteins. The major advantage of these experiments is that they can measure the FRET efficiencies of the DSE under solution conditions where the native state is stable. The average denaturant-dependent FRET efficiency $\langle E\rangle$ has been used to infer the global properties of the polypeptide chain in the DSE as the external conditions are altered. The properties of the DSE are inferred from $\langle E\rangle$ by assuming a polymer model for the DSE, from which the root mean squared distance between two dyes attached at residues i and j along the protein sequence ($R_{ij}=\langle|\mathbf{r}_i-\mathbf{r}_j|\rangle$), the distribution of the end-to-end distance $P(R)$ (where $R=|\mathbf{r}_N-\mathbf{r}_0|$), the root mean squared end-to-end distance ($R_{ee}=\langle\mathbf{R}^2\rangle^{1/2}$), the root mean squared radius of gyration ($R_g=\langle\mathbf{R}_g^2\rangle^{1/2}$), and the persistence length (l_p) of the denatured protein[6–15] can be calculated.

In FRET experiments, donor (D) and acceptor (A) dyes are attached at two locations along the protein sequence,[4,16] and hence can only provide information about correlations

[a)] Author to whom correspondence should be addressed. Tel.: 301-405-4803. FAX: 301-314-9404. Electronic mail: thirum@umd.edu.

TABLE I. Polymer models and their properties.

	Property		
Polymer model	End-to-end distribution $P(R)$ [a]	Radius of gyration R_g	Persistence length l_p
Gaussian	$4\pi R^2\left(\frac{3}{2\pi Na^2}\right)^{3/2}\exp\left(\frac{-3R^2}{2Na^2}\right)$	$a\sqrt{N/6}$	$\frac{Na^2}{2L}=\frac{a}{2}$
WLC[b]	$\frac{4\pi R^2 C_1}{L(1-(R/L)^2)^{9/2}}\exp\left(\frac{-3L}{4l_p(1-(R/L)^2)}\right)$	$\frac{L}{6C_2}+\frac{1}{4C_2^2}+\frac{1}{4LC_2^3}-\frac{1-\exp(-L/l_p)}{8C_2^4L^2}$	$R_{ee}^2=2l_pL-2l_p^2-2l_p^2\exp\left(-\frac{L}{l_p}\right)$ [c]
Self-avoiding polymer[d]	$\frac{a}{R_{ee}}\left(\frac{R}{R_{ee}}\right)^{2+\theta}\exp\left(-b\left(\frac{R}{R_{ee}}\right)^{\delta}\right)$	N/A	N/A

[a] The average end-to-end distance $R_{ee}=(\int R^2P(R)dR)^{1/2}$.

[b] L and l_p are the contour length and persistence length, respectively. $C_1=(\pi^{3/2}e^{-\alpha}\alpha^{-3/2}(1+3\alpha^{-1}+\frac{15}{4}\alpha^{-2}))^{-1}$, where $\alpha=3L/(4l_p)$. $C_2=1/(2l_p)$.

[c] Using the simulated $\langle R^2\rangle$, l_p was solved for numerically using this equation.

[d] θ and δ equal to 0.3 and 2.5, respectively. The constants a and b are determined by solving the integrals of the zeroth and second moment of $\int P(R)dr=\int R^2P(R)dr=1$, resulting in values of $a=3.678\,53$ and $b=1.231\,52$.

between them. The efficiency of energy transfer E between the D and A is equal to $(1+r^6/R_0^6)^{-1}$, where r is the distance between the dyes, and R_0 is the dye-dependent Förster distance.[4,16] Because of conformational fluctuations, there is a distribution of r, $P(r)$, which depends on external conditions such as the temperature and denaturant concentration. As a result, the average FRET efficiency $\langle E\rangle$ is given by

$$\langle E\rangle=\int_0^{\infty}(1+r^6/R_0^6)^{-1}P(r)dr \qquad (1)$$

under most experimental conditions due to the central limit theorem.[17] If the dyes are attached to the ends of the chain, then $P(r)=P(R)$. Even if $\langle E\rangle$ is known accurately, the extraction of $P(R)$ from the integral equation [Eq. (1)] is fraught with numerical instabilities. In experimental applications to biopolymers, a functional form for $P(r)$ is assumed in order to satisfy the equality in Eq. (1). The form of $P(r)$ is based off of a particular polymer model which depends only on a single parameter (see Table I): the Gaussian chain (dependent on the Kuhn length a), the wormlike chain (WLC) (dependent on the persistence length l_p), and the self-avoiding walk (SAW) (dependent on the average end-to-end distance R_{ee}). For the chosen polymer model meant to represent the biopolymer of interest, the free parameter (a, l_p, or R_{ee}) is determined numerically to satisfy Eq. (1). Using this method (referred to as the "standard procedure" in this article), several researchers have estimated R_g and l_p as a function of the external conditions for protein L,[11,14] cold shock protein (CspTm),[13] and Rnase H.[16] The justification for using homopolymer models to analyze FRET data comes from the anecdotal comparison of the R_g measured using x-ray scattering experiments and the extracted R_g from analysis of Eq. (1).[4]

Here, we study an analytically solvable generalized Rouse model (GRM)[18] and the Molecular Transfer Model (MTM) for protein L[19] to assess the accuracy of using polymer models to solve Eq. (1). In the GRM, two monomers that are not covalently linked interact through a harmonic potential that is truncated at a distance c. The presence of the additional length scale, c, which reflects the interaction between nonbonded beads, results in the formation of an ordered state as the temperature (T) is varied. A more detailed discussion of these models can be found in Sec. IV. For the GRM, $P(R)$ can be analytically calculated, and hence the reliability of the standard procedure to solve Eq. (1) can be unambiguously established. We find that the accuracy of the polymer models in extracting the exact values in the GRM depends on the location of the monomers that are constrained by the harmonic interaction. Using coarse-grained simulations of protein L, we show that the error between the exact quantity and that inferred using the standard procedure depends on the property of interest. For example, the inferred end-to-end distribution $P(R)$ is in qualitative, but not quantitative agreement with the exact $P(R)$ distribution obtained from accurate simulations. In general, the DSE of protein L is better characterized by the SAW polymer model than the Gaussian chain model.

We propose that the accuracy of the popular Gaussian model can be assessed by measuring $\langle E\rangle$ with dyes attached at multiple sites in a protein.[13,20,21] If the DSE can be described by a Gaussian chain, then the parameters extracted by attaching the dyes at position i and j can be used to predict $\langle E\rangle$ for dyes at other points. The proposed self-consistency test shows that the Gaussian model only qualitatively accounts for the experimental data of CspTm, simulation results for protein L, and the exact analysis of the GRM.

II. RESULTS AND DISCUSSION

We present the results in three sections. In Secs. II A and II B we examine the accuracy of the standard procedure (described in Sec. I) in accurately inferring the properties of the denatured state of the GRM and protein L models. Section II C presents results of the Gaussian self-consistency (GSC) test applied to these models. We also analyze experimental data for CspTm to assess the extent to which the DSE deviates from a Gaussian chain.

A. GRM

The GRM is a simple modification of the Gaussian chain with N bonds and Kuhn length a_0, which includes a single, noncovalent bond between two monomers at positions s_1 and

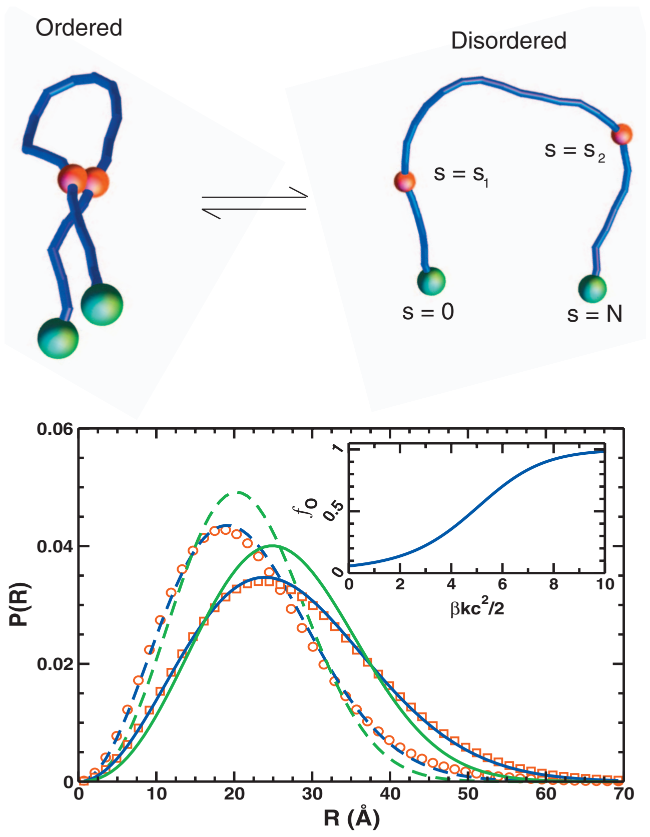

FIG. 1. (Color) Top figures show a schematic sketch of the GRM, with the donor and acceptor at the end points, represented by the green spheres, and the interacting monomers at s_1 and s_2 represented by the red spheres. In the ordered configuration, the monomers at s_1 and s_2 are tightly bound. The bottom figure shows the exact and the inferred end-to-end distribution functions $P(r)$ for interior interactions ($\Delta s=31$). The blue lines correspond to the Gaussian chain model, light green lines to the SAW, and the symbols to the exact GRM distribution. Dashed lines and red circles are for $\beta\kappa=6.6$, while solid lines and red squares correspond to $\beta\kappa=2$. In the inset we show the fraction of ordered states as a function of $\beta\kappa$. Note that 75% of the structures are ordered at $\beta\kappa=6.6$, yet the inferred Gaussian $P(r)$ is in excellent agreement with the exact result.

s_2 (Fig. 1). The monomers at s_1 and s_2 interact with a truncated harmonic potential with spring constant k, with strength $\kappa=kc^2/2$, where c is the distance at which the interaction vanishes [Eq. (4)]. The GRM minimally represents a two-state system, with a clear demarcation between ordered [with $|\mathbf{r}(s_2)-\mathbf{r}(s_1)|\le c$] and disordered [with $|\mathbf{r}(s_2)-\mathbf{r}(s_1)|>c$] states. Unlike other polymer models (see Table I), which are characterized by a single length scale, the GRM is described by a_0 and the energy scale κ. For $\beta\kappa\to 0$ (the high temperature limit, where $\beta=1/k_BT$), the simple Gaussian chain is recovered (see Sec. IV for details). By varying $\beta\kappa$, a disorder → order transition can be induced (see Fig. 1). The presence of the interaction between monomers s_1 and s_2 approximately mimics persistence of structure in the DSE of proteins. If the fraction of ordered states, f_O, exceeds 0.5 (Fig. 1 inset), we assume that the residual structure is present with high probability. The exact analysis of the GRM when $|\mathbf{r}(s_2)-\mathbf{r}(s_1)|\le c$ allows us to examine the effect of structure in the DSE on the global properties of unfolded states.

Because $\langle E\rangle$ can be calculated exactly for the GRM [see Eq. (5)], it can be used to quantitatively study the accuracy of solving Eq. (1) using the standard procedure.[6,10,11,13,14] Given the best fit for the Gaussian chain (Kuhn length a), WLC (persistence length l_p), and SAW (average end-to-end distance R_{ee}), as described in Table I, many quantities of interest can be inferred [$P(R)$ or R_g, for example], and compared to the exact results for the GRM. The extent to which the exact and inferred properties deviate, due to the additional single energy scale in the GRM, is an indication of the accuracy of the standard procedure used to analyze Eq. (1).

1. $P(R)$ is accurately inferred using the Gaussian polymer model

If the interacting monomers are located near the end points of the chain, the end-to-end distribution function is bimodal, with a clear distinction between the ordered and disordered regions.[18] However, if the monomers s_1 and s_2 are in the interior of the chain, the two-state behavior is obscured because the distribution function becomes unimodal. In Fig. 1, we show the exact and inferred $P(R)$ functions for a chain with $N=63$, $a_0=3.8$ Å, $c=2a_0$, and $|s_2-s_1|=(N-1)/2=31$. We take the Förster distance [Eq. (1)] $R_0=23$ Å $\lesssim\langle\mathbf{R}^2\rangle^{1/2}_{\kappa=0}$ for the GRM. The distributions are unimodal for both weakly ($\beta\kappa=2$) and strongly ($\beta\kappa=6.6$) interacting monomers.

The strength of the interaction is most clearly captured with the fraction of conformations in the ordered state, f_O, with $f_O=0.25$ for the weakly interacting chain and $f_O=0.75$ for the strongly interacting chain (inset of Fig. 1). The inferred Gaussian distribution functions are in excellent agreement with the exact result. Because of the underlying Gaussian Hamiltonian in the GRM, the rather poor agreement in the inferred SAW distribution seen in Fig. 1 is to be expected. We also note that the GRM is inherently flexible so that the WLC and Gaussian chains produce virtually identical distributions.

2. The accuracy of the inferred R_g depends on the location of the interaction

The two-state nature of the GRM is obscured by the relatively long unstructured regions of the chain, similar to the effect seen in laser optical tweezers experiments with flexible handles.[18] As a result, $P(R)$ is well represented by a Gaussian chain, with a smaller inferred Kuhn length, $a\le a_0$ (Fig. 2). For large $\beta\kappa$, where the ordered state is predominantly occupied and $\mathbf{r}(s_2)\approx\mathbf{r}(s_1)$, the end-to-end distribution function is well approximated by a Gaussian chain with $N^*=N-\Delta s$ bonds. Consequently, the single length scale for the Gaussian chain decreases to $a\sim a_0\sqrt{1-\Delta s/N}\approx 0.71a_0$ for large values of $\beta\kappa$ (Fig. 2).

Because the two-state nature of the chain is obscured for certain values of $|s_2-s_1|$, the Gaussian chain gives an excellent approximation to the end-to-end distribution function. However, the radius of gyration R_g is not as accurately obtained using the Gaussian chain model, as shown in Fig. 3. The exact R_g for the GRM reflects both the length scale a_0 and the energy scale $\beta\kappa$, which can not be fully described by the single inferred length scale a in the Gaussian chain. For the GRM, R_g depends not only on the separation between the monomers Δs, but also explicitly on s_1 (i.e., where the inter-

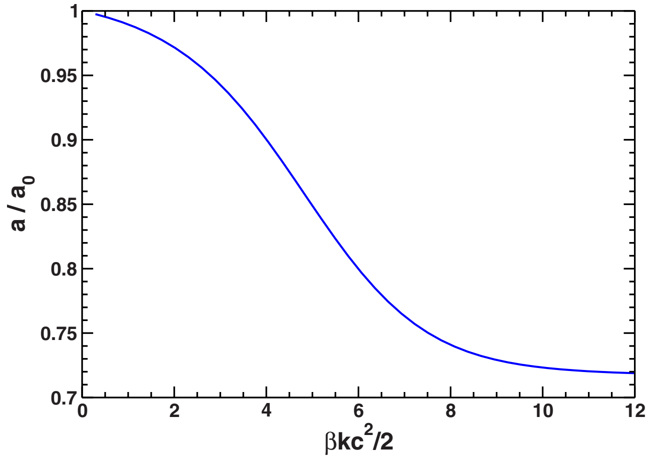

FIG. 2. (Color online) The inferred Kuhn length a as a function of $\beta\kappa$ for the GRM. R_{ee} monotonically decreases a function of the interaction strength, leading to the decrease in a/a_0. The Kuhn length a reaches its limiting value of $a \approx a_0\sqrt{1-\Delta s/N}$ when $f_O \approx 1$.

action is along the chain; see Fig. 3 and Sec. IV), which can not be captured by the Gaussian chain. If the interacting monomers are in the middle of the chain $[s_1=(N+1)/4=16$ and $\Delta s=31]$, the inferred R_g is in excellent agreement with the exact result (Fig. 3). The relative error in R_g (the difference between the inferred and exact values, divided by the exact value) is no less than -2%. However, for interactions near the end point of the chain, with $s_1=0$ and the same $\Delta s=31$, the relative error between the inferred and exact values of R_g is $\sim -14\%$. The large errors arise because the radius of gyration depends on the behavior of all of the monomers so that the energy scale $\beta\kappa$ plays a much larger role in the determination of R_g than R_{ee}.

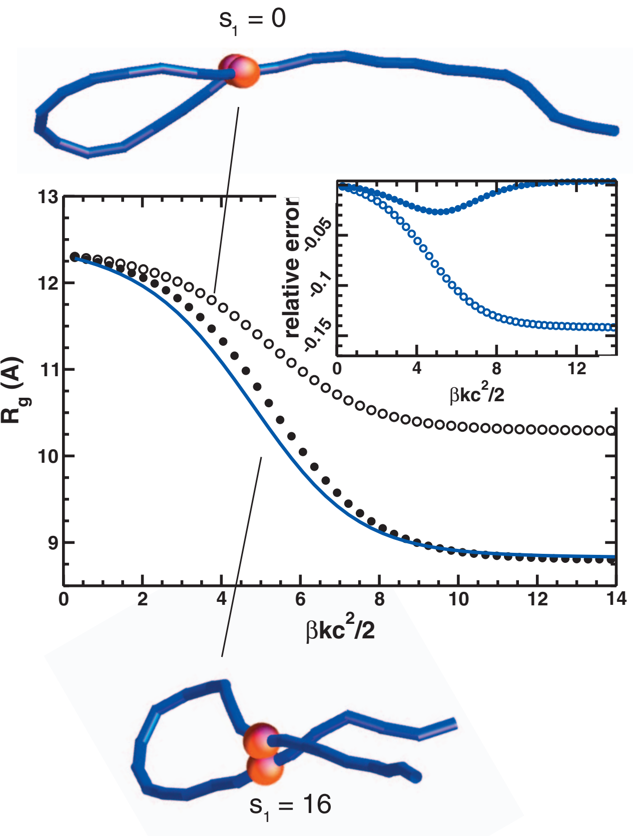

FIG. 3. (Color) Comparison of the exact (symbols) and inferred (blue line) values of the radius of gyration (R_g) as a function of $\beta\kappa$ for $\Delta s=31$. Shown are R_g's for the GRM with $s_1=0$ (open symbols) and $s_1=16$ (filled symbols) for $N=63$. The structures in the ordered state are shown schematically. The R_g obtained using the standard procedure is independent of s_1, while the exact result is not. The inset shows the relative errors between the inferred and exact values of R_g.

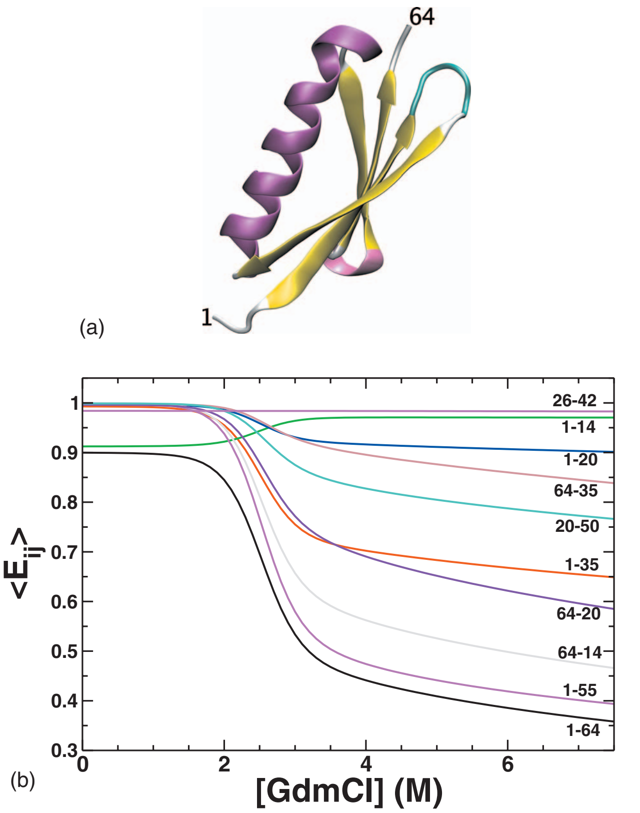

FIG. 4. (Color) (a) A secondary structure representation of protein L in its native state. Starting from the N-terminus, the residues are numbered 1–64. (b) The average FRET efficiency between the various (i,j) residue pairs in protein L vs GdmCl concentration. The $\langle E_{ij}\rangle$ values, computed using MTM simulations, for each (i,j) pair is indicated by the two numbers next to each line. For example, the numbers "1–64" beneath the black line indicates that $i=1$ and $j=64$. The solid black line (lowest values of $\langle E\rangle$) is computed for the dyes at the end points.

B. MTM for protein L

Protein L is a 64 residue protein [Fig. 4(a)] whose folding has been studied by a variety of methods.[11,14,22–24] More recently, single molecule FRET experiments have been used to probe changes in the DSE as the concentration of GdmCl is increased from 0 to $7M$.[11,14] From the measured GdmCl-dependent $\langle E\rangle$, the properties of the DSE, such as R_{ee}, $P(R)$, and R_g, were extracted by solving Eq. (1), and assuming a Gaussian chain $P(R)$.[11,14] To further determine the accuracy of polymer models in the analysis of $\langle E\rangle$, we use simulations of protein L in the same range of the concentration of denaturant, [C], as used in experiments.[6,9]

1. The average end-to-end distance is accurately inferred from FRET data

In a previous study,[19] we showed that the predictions based on MTM simulations for protein L are in excellent

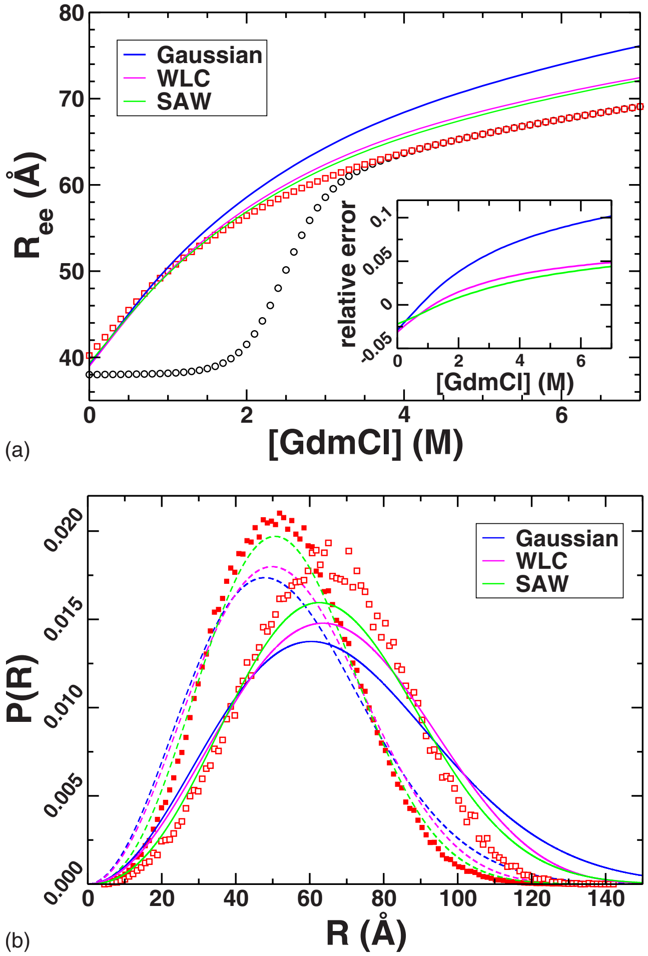

FIG. 5. (Color online) (a) The root mean squared end-to-end distance (R_{ee}) as a function of GdmCl concentration for protein L. The average R_{ee} (black circles) and R for the subpopulation of the DSE [(red) squares] from simulations are shown. The values of R_{ee} inferred by solving Eq. (1) by the standard procedure using the Gaussian chain, WLC, and SAW polymer models are shown for comparison as the top, middle, and bottom solid lines, respectively. The inset shows the relative errors between the exact and the values inferred using the FRET efficiency for R_{ee} vs GdmCl concentration. The top, middle, and bottom lines correspond to the Gaussian chain, WLC, and SAW polymer models, respectively. (b) Simulation results of the denatured state end-to-end distance distribution ($P(R)$) at 2.4M GdmCl [solid (red) squares] and 6M GdmCl [open (red) squares] and T=327.8 K are compared with $P(R)$s using the Gaussian chain, WLC, and SAW polymer models are also shown at 2.4M GdmCl (dashed lines) and 6M GdmCl (solid lines). The top, middle, and bottom lines correspond to the SAW, WLC, and Gaussian chain polymer models, respectively.

agreement with experiments. From the calculated $\langle E \rangle$ with the dyes at the end points [solid black line in Fig. 4(b)], which is in quantitative agreement with experimental measurements,[19] we determine the model parameter R_{ee} or l_p by assuming that the exact $P(R)$ can be approximated by the three polymer models in Table I. Comparison of the exact value of R_{ee} to the inferred value R_F, obtained using the simulation results for $\langle E \rangle$, shows good agreement for all three polymer models [Fig. 5(a)]. There are deviations between R_{ee} and R_F at [C]>C_m, the midpoint of the folding transition. The maximum relative error [see inset of Fig. 5(A)] we observe is about 10% at the highest concentration of GdmCl. The SAW model provides the most accurate estimate of R_{ee} at GdmCl concentrations above C_m, with a relative error ≤ 0.05, and the Gaussian model gives the least accurate values, with a relative error ≤ 0.10 [Fig. 5(a)]. Due to the relevance of excluded volume interaction in the DSE of real proteins, the better agreement using the SAW is to be expected.

2. Polymer models do not give quantitative agreement with the exact *P*(*R*)

The inferred distribution functions, $P_F(R)$'s, obtained by the standard procedure (as described in the introduction) at [C]=2M and 6M GdmCl differ from the exact results [Fig. 5(b)]. Surprisingly, the agreement between $P(R)$ and $P_F(R)$ is worse at higher [C]. The range of R explored and the width of the exact distribution are less than predicted by the polymer models. The Gaussian chain and the SAW models account only for chain entropy, while the WLC only models the bending energy of the protein. However, in protein L (and in other proteins) intramolecular attractions are still present even when [C]=6M>C_m. As a result, the range of R explored in the protein L simulations is expected to be less than in these polymer models. Only at $[\mathrm{C}]/C_m \gg 1$ and/or at high T are proteins expected to be described by Flory random coils. Our results show that although it is possible to use models that can give a single quantity correctly (R_{ee}, for example), the distribution functions are less accurate. The results in Fig. 5(b) show that $P(R)$, inferred from the polymer models, agrees only qualitatively with the exact $P(R)$, with the SAW model being the most accurate [Fig. 5(b)]. While the MTM will not perfectly reproduce all of the fine details of protein L under all situations, we expect it to produce more realistic results than idealized polymer models, which have no specific intrachain interactions.

3. Inferred R_g and l_p differ significantly from the exact values

The solution of Eq. (1) using a Gaussian chain or WLC model yields a and l_p, from which R_g can be analytically calculated (Table I). Figures 6(a) and 6(b), which compare the FRET inferred R_g and l_p with the corresponding values obtained using MTM simulations, show that the relative errors are substantial. At high [C] values the R_g^F deviates from R_g by nearly 25% if the Gaussian chain model is used [Fig. 6(a)]. The value of $R_g \approx 26$ Å at [C]=8 M while R_g^F using the Gaussian chain model is ≈31 Å. In order to obtain reliable estimates of R_g, an accurate calculation of the distance distribution between all the heavy atoms in a protein is needed. Therefore, it is reasonable to expect that errors in the inferred $P(R)$ are propagated, leading to a poor estimate of internal distances, thus resulting in a larger error in R_g. A similar inference can be drawn about the persistence length obtained using polymer models [Fig. 6(b)]. Plotting l_p^F as a function of [C] [Fig. 6(b)], against $l_p = R_{ee}/2L$, shows that l_p is overestimated at concentrations above 1M GdmCl, with the error increasing as [C] increases. The error is less when the Gaussian chain model is used.

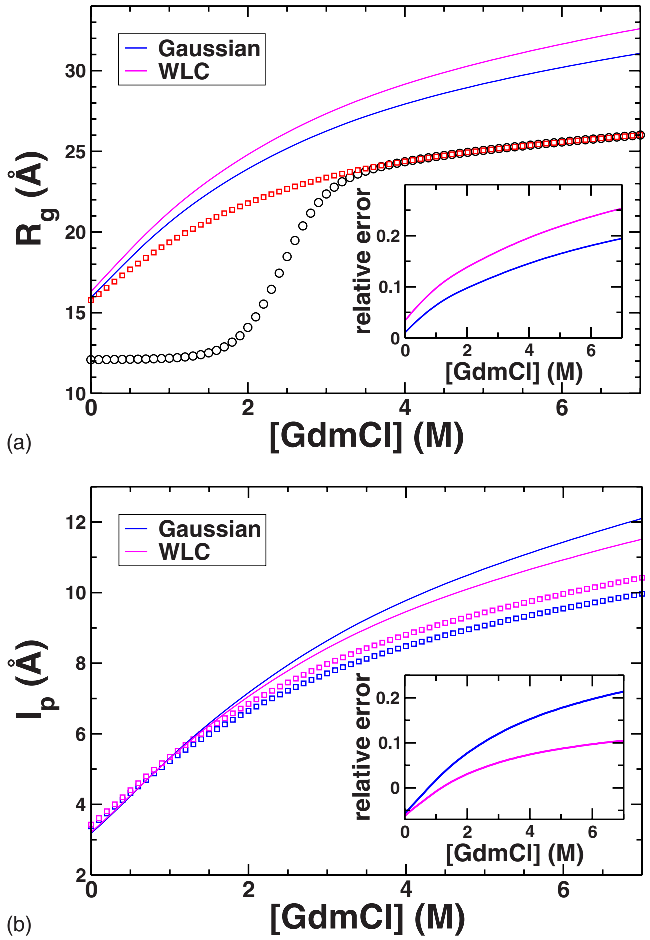

FIG. 6. (Color online) (a) Comparison of R_g from direct simulations of protein L and that obtained by solving Eq. (1) using the Gaussian chain and WLC polymer models. The top line (magenta) shows the WLC fit, the bottom line (blue) shows the Gaussian fit, squares (red) show the DSE R_g from the simulation, and black circles show the average simulated R_g. The inset shows the relative errors as a function of GdmCl concentration; top and bottom lines correspond to the WLC and Gaussian chain polymer models, respectively. (b) Same as (a) except the figure is for l_p. Top and bottom lines correspond to the inferred l_p using the WLC and Gaussian chain polymer models, respectively. Top and bottom sets of squares correspond to a direct analysis of the simulations using the WLC and Gaussian chain polymer models, respectively.

C. Gaussian self-consistency test shows the DSE is non-Gaussian

The extent to which the Gaussian chain accurately describes the ensemble of conformations that are sampled at different values of the external conditions (temperature or denaturants) can be assessed by performing a self-consistency test. A property of a Gaussian chain is that if the average root mean square distance, R_{ij}, between two monomers i and j is known then R_{kl}, the distance between any other pair monomers k and l, can be computed using

$$R_{kl}=\sqrt{\frac{|k-l|}{|i-j|}}R_{ij}. \tag{2}$$

Thus, if the conformations of a protein (or a polymer) can be modeled as a Gaussian chain, then R_{ij} inferred from the FRET efficiency $\langle E_{ij}\rangle$ should accurately predict R_{kl} and the FRET efficiency $\langle E_{kl}\rangle$, if the dyes were to be placed at monomers k and l. We refer to this criterion as the GSC test, and the extent to which the predicted R_{kl} from Eq. (2) deviates from the exact R_{kl} reflects deviations from the Gaussian model description of the DSE.

1. GSC test for the GRM

For the GRM, with a nonbonded interaction between monomers s_1 and s_2, we calculate $\langle E_{ij}\rangle$ using Eq. (8) with j fixed at 0 and for $i=20$, 40, and 60. Using the exact results for $\langle E_{ij}\rangle$, the values of R_{ij} are inferred assuming that $P(r)$ is a Gaussian chain. From the inferred R_{ij} the values of $\langle E_{kl}\rangle$ and R_{kl} can be calculated using Eqs. (1) and (2), respectively. We note that since $R_{kl}/R_{ij}=\sqrt{|k-l|/|i-j|}$ [Eq. (2)] for any pair (k,l) using the Gaussian chain model, the prediction of the Gaussian chain will be independent of the particular choices of k and l, as long as their difference is held constant. We first apply the GSC test to a GRM in which $f_O\approx 0.75$ due to a favorable interaction between monomers $s_1=16$ and $s_2=47$. There are discrepancies between the values of the Gaussian inferred (R_{kl}^G) and exact R_{kl} distances, as well as the inferred ($\langle E_{kl}^G\rangle$) and exact $\langle E_{ij}\rangle$ efficiencies when a Gaussian model is used (Fig. 7). The relative errors in the predicted values of the FRET efficiency and the interdye distances can be as large as 30%–40%, depending on the choice of i and j (see insets in Fig. 7). We note that the relative error in the end-to-end distance is small for dyes near the end points [the green line in Fig. 7(b)], in agreement with the results shown in Fig. 1. The errors decrease as f_O decreases, with a maximum error of 20% when $f_O=0.5$, and 10% when $f_O=0.25$ (data not shown). By construction, the GRM is a Gaussian chain when $f_O=0$ and therefore the relative errors will vanish at sufficiently small $\beta\kappa$ (data not shown). These results show that even for the GRM, with only one nonbonded interaction in an otherwise Gaussian chain, its DSE cannot be accurately described using a Gaussian chain model. Thus, even if the overall end-to-end distribution $P(r)$ for the GRM is well approximated as a Gaussian (as seen in Fig. 1), the internal R_{kl} monomer pair distances can deviate from predictions of the Gaussian chain model.

2. GSC test for protein L

We apply the GSC test to our simulations of protein L at GdmCl concentrations of $[C]=2.0M$ (below $C_m=2.4M$) and $[C]=7.5M$ (well above C_m). While our simulations allow us to compute the DSE $\langle E_{ij}\rangle$ for all possible (i,j) pairs, we examine only a subset of $\langle E_{ij}\rangle$ as a function of GdmCl concentration [Fig. 4(b)]. By choosing multiple j values for the same value of i, we can determine whether distant residues along the backbone are close together spatially, which may offer insights into three-point correlations in denatured states. We note that all values of $\langle E_{ij}\rangle$ in Fig. 4 are monotonically decreasing, except for the (1,14) pair. This is due to the fact that the native state has a beta strand between these two residues; as the protein denatures, they come closer together, increasing the FRET efficiency. We use these values for $\langle E_{ij}\rangle$ in the GSC test. The results are shown in Figs. 8(a) and 8(b). Relative errors in $\langle E_{kl}\rangle$ as large as 36% at $2.0M$ GdmCl and 50% at $7.5M$ GdmCl are found, with the lowest errors generally seen for residues close to one another along the backbone, in agreement with the results from the GRM [Fig. 7(a) inset]. In addition, the number of data points that underestimate $\langle E_{kl}\rangle$ increases as [C] is changed from 7.5 to 2.0M for $|k-l|<20$. Despite these differences, the gross features in Figs. 8(a) and 8(b) are concentration independent. Because the error does not vanish for all (k,l) pairs [Figs. 8(a) and 8(b)], we conclude that the DSE of protein L cannot be modeled as a Gaussian chain.

3. GSC test for CspTm

In an interesting single molecule experiment, Hoffmann *et al.*[13] measured FRET efficiencies by attaching donor and acceptor dyes to pairs of residues at five different locations of a CspTm. They analyzed the data by assuming that the DSE properties can be mimicked using a Gaussian chain model. We used the GSC test to predict $\langle E_{kl}\rangle$ for dyes separated by $|k-l|$ along the sequence using the experimentally measured values $\langle E_{ij}\rangle$.

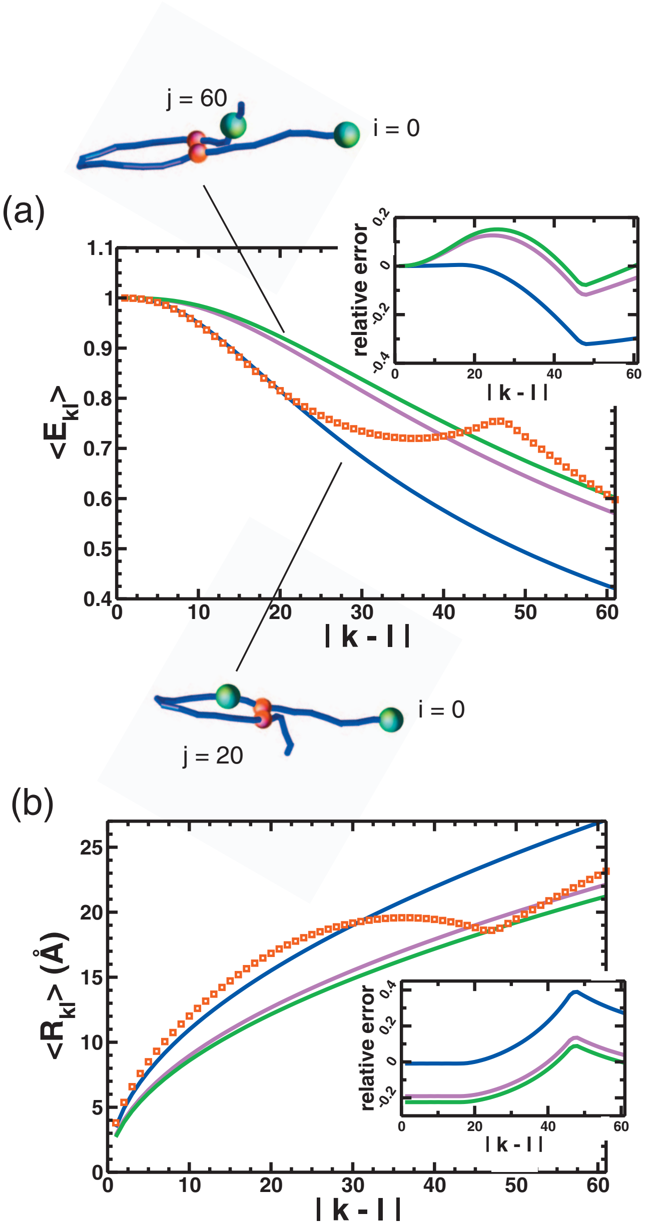

FIG. 7. (Color) GSC test using (a) the FRET efficiency and (b) the average end-to-end distance for the GRM with f_O=0.75 and interaction sites at s_1=16 and s_2=47. In both (a) and (b) the solid lines are the inferred properties and the open symbols are the exact values. In both (a) and (b), j=0 and the blue, magenta, and green lines correspond to a dye at i=20, 40, and 60, respectively. The insets show the relative error for $\langle E_{kl}\rangle$ and R_{kl}. Note that the relative error would be zero if the Gaussian chain accurately modeled the GRM.

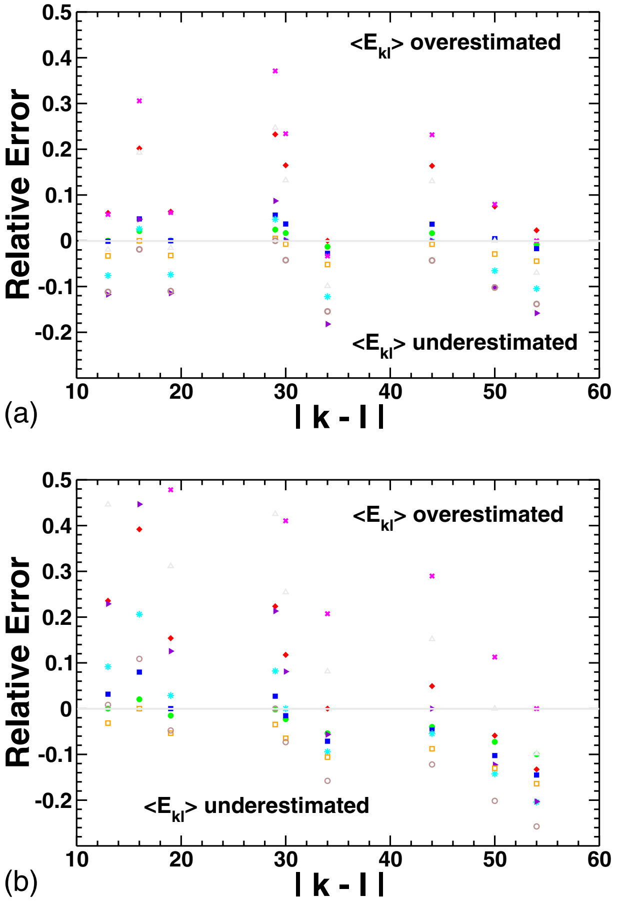

FIG. 8. (Color online) The Gaussian self-consistency test applied to simulated DSE $\langle E_{ij}\rangle$ data of protein L using the (i,j) pairs listed in Fig. 4(b). Shown are the relative errors at (a) 2.0M GdmCl and (b) 7.5M GdmCl. In both (a) and (b), solid (green) circles correspond to $|i-j|=13$, open (orange) squares to $|i-j|=16$, solid (blue) squares to $|i-j|=19$, open (brown) circles to $|i-j|=29$, asterisks (cyan) to $|i-j|=30$, diamonds (red) to $|i-j|=34$, solid (violet) triangles to $|i-j|=44$, open (gray) triangles to $|i-j|=50$, and crosses (magenta) to $|i-j|=54$. Each symbol corresponds to a line in Fig. 4(b), with the colors of the symbols corresponding to the colors of the line, except for the 1–64 pair (not shown here).

The relative error in $\langle E_{kl}\rangle$ [Eq. (2)] should be zero if CspTm can be accurately modeled as a Gaussian chain. However, there are significant deviations (up to 17%) between the predicted and experimental values (Fig. 9). The relative error is fairly insensitive to the denaturant concentration [compare Figs. 9(a) and 9(b)]. It is interesting to note that the trends in Fig. 9 are qualitatively similar to the relative errors in the GRM at $f_O>0$. Based on these observations we conclude tentatively that whenever the DSE is ordered to some extent (i.e., when there is persistent residual structure) then we expect deviations from a homopolymer description of the DSE of proteins. At the very least, the GSC test should be routinely used to assess errors in the modeling of the DSE as a Gaussian chain.

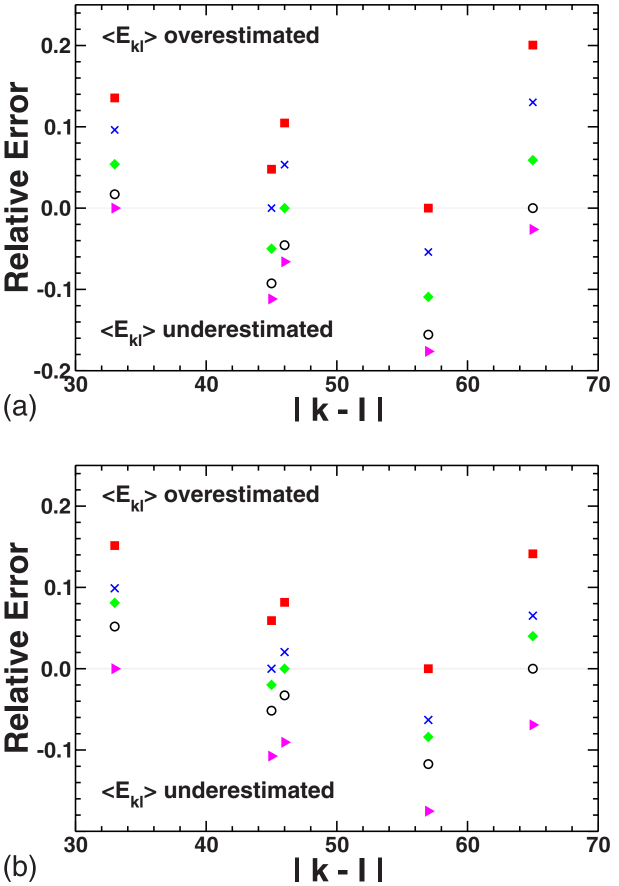

FIG. 9. (Color online) The GSC test using experimental data from CspTm. One dye was placed at one end point, and the location of the other was varied. We show relative error of the predicted $\langle E \rangle$, using Eqs. (1) and (2), versus the distance between the dyes ($|k-l|$) for [C]=2M (a) and 5M (b). In both (a) and (b), triangles correspond to $|i-j|=33$, x's to $|i-j|=45$, diamonds to $|i-j|=46$, squares to $|i-j|=57$, and circles to $|i-j|=65$. The trends in Figs. 7 and 8 are similar.

III. CONCLUSIONS

In order to assess the accuracy of polymer models to infer the properties of the DSE of proteins from measurement of FRET efficiencies, we studied two models for which accurate calculations of all the equilibrium properties can be carried out. Introduction of a nonbonded interaction between two monomers in a Gaussian chain (the GRM) leads to an disorder-order transition as the temperature is lowered. The presence of "residual structure" in the GRM allows us to clarify its role in the use of the Gaussian chain model to fit the accurately calculated FRET efficiency. Similarly, we have used the MTM model for protein L to calculate precisely the denaturant-dependent $\langle E \rangle$ from which we extracted the global properties of the DSE by solving Eq. (1) using the $P(R)$'s for the polymer models in Table I. Quantitative comparison of the exact values of a number of properties of the DSE (obtained analytically for the GRM and accurately using simulations for protein L) and the values inferred from $\langle E \rangle$ has allowed us to assess the accuracy with which polymer models can be used to analyze the experimental data. The major findings and implications of our study are listed below.

(1) The polymer models, in conjunction with the measured $\langle E \rangle$, can accurately infer values of R_{ee}, the average end-to-end distance. However, $P(R)$, l_p, and R_g are not quantitatively reproduced. For the GRM, R_g is underestimated, whereas it is overestimated for protein L. The simulations show that the absolute value of the relative error in the inferred R_g can be nearly 25% at elevated GdmCl concentration.

(2) We propose a simple self-consistency test to determine the ability of the Gaussian chain model to correctly infer the properties of the DSE of a polymer. Because the Gaussian chain depends only on a single length scale, the FRET efficiency can be predicted for varying dye positions once $\langle E \rangle$ is accurately known for one set of dye positions. The GSC test shows that neither the GRM, simulations of protein L, nor experimental data on CspTm can be accurately modeled using the Gaussian chain. The relative errors between the exact and predicted FRET efficiencies can be as high as 50%. For the GRM, we find that the variation in the FRET efficiency as a function of the dye position changes abruptly if one dye is placed near an interacting monomer. Taken together these findings suggest that it is possible to infer the structured regions in the DSE by systematically varying the location of the dyes. This is due to the fact that the FRET efficiency is perfectly monotonic using the Gaussian chain model. An experiment that shows nonmonotonic behavior in $\langle E_{ij} \rangle$ as the dye positions i and j are varied is a clear signal of non-Gaussian behavior, and sharp changes in the FRET efficiency as a function of $|i-j|$ may indicate strongly interacting sites [see Fig. 7(a)].

(3) The properties of the DSE inferred from Eq. (1) become increasingly more accurate as [C] decreases. At a first glance this finding may be surprising, especially considering that stabilizing intrapeptide interactions are expected to be weakened at high GdmCl concentrations [C], and therefore the protein should be more "polymerlike." The range of R-values sampled at low [C] is much smaller than at high [C]. Protein L swells as [C] is increased, as a consequence of the increase in the solvent quality. It is possible that [C]$\approx$2.4M might be close to a Θ-solvent (favorable intrapeptide and solvent-peptide interactions are almost neutralized) so that $P(R)$ can be approximated by a polymer model. The inaccuracy of polymer models in describing $P(R)$ at [C]=6M suggests that only at much higher concentrations does protein L behave as a random coil. In other words, T=327.8 K and [C]=6M is not an athermal (good) solvent.

(4) It is somewhat surprising that polymer models, which do not have side chains or any preferred interactions between the beads, are qualitatively correct in characterizing the DSE of proteins with complex intramolecular interactions. In addition, even [C]=6M GdmCl is not an athermal solvent, suggesting that at lower [C] values the aqueous denaturant may be closer to a Θ-solvent. A consequence of this observation is that for many globular proteins, the extent of collapse may not be significant, resulting in the nearness of the concen-

trations at which collapse and folding transitions occur, as shown by Camacho and Thirumalai[25] some time ago. We suggest that only by exploring the changes in the conformations of polypeptide chains over a wide range of temperature and denaturant concentrations can one link the variations of the DSE properties (compaction) and folding (acquisition of a specific structure).

IV. THEORY AND COMPUTATIONAL METHODS

A. GRM model

In order to understand the effect of a single noncovalent interaction between two monomers along a chain, we consider a Gaussian chain with Kuhn length a_0 and N bonds, with a harmonic attraction between monomers $s_1 \leq s_2$, which is cutoff at a distance c. The Hamiltonian for the GRM is

$$\beta H = \frac{3}{2a^2}\int_0^N ds\dot{\mathbf{r}}^2(s) + \beta V[\mathbf{r}(s_2) - \mathbf{r}(s_1)], \tag{3}$$

$$\beta V[\mathbf{r}] = \begin{cases} k\mathbf{r}^2/2, & |\mathbf{r}| < c \\ kc^2/2, & |\mathbf{r}| \geq c, \end{cases} \tag{4}$$

where k is the spring constant that constrains $\mathbf{r}(s_2)-\mathbf{r}(s_1)$ to a harmonic well. The Hamiltonian in Eq. (3) allows the exact determination of many quantities of interest. Defining $\mathbf{x} = \mathbf{r}(s_2)-\mathbf{r}(s_1)$ and $\Delta s = s_2 - s_1$, we can determine most averages of interest for the GRM using

$$\langle \cdots \rangle = \frac{\int d^3\mathbf{x}d^3\mathbf{r}_N(\cdots)G(\mathbf{x},\mathbf{r}_N;\Delta s,N)}{\int d^3\mathbf{x}d^3\mathbf{r}_N G(\mathbf{x},\mathbf{r}_N;\Delta s,N)}, \tag{5}$$

$$G(\mathbf{x},\mathbf{r}_N;\Delta s,N) = \exp\left(-\frac{3\mathbf{x}^2}{2\Delta s a^2} - \frac{3(\mathbf{r}_N-\mathbf{x})^2}{2(N-\Delta s)a^2} - \beta V[\mathbf{x}]\right). \tag{6}$$

B. C$_\alpha$-SCM protein model and GdmCl denaturation

We use the coarse-grained C$_\alpha$-side chain model (C$_\alpha$-SCM) to model protein L (for details, see the supporting information in Ref. 19). In the C$_\alpha$-SCM each residue in the polypeptide chain is represented using two interaction sites, one that is centered on the α-carbon atom and another that is located at the center of mass of the side chain.[26] Langevin dynamics simulations[27] are carried out in the underdamped limit at zero molar guanidinium chloride. Simulation details are given in.[19]

We model the denaturation of protein L by GdmCl using the MTM.[19] MTM combines simulations at zero molar GdmCl with experimentally measured transfer free energies, using a reweighing method[28–30] to predict the equilibrium properties of proteins at any GdmCl concentration of interest.

V. ANALYSIS

A. GRM

The average squared end-to-end distance can be computed directly from Eq. (5), using $\langle \mathbf{R}_{ee}^2\rangle = Na_0^2 + (\langle \mathbf{x}^2\rangle - \Delta s a_0^2)$. The exact expression for $\langle \mathbf{x}^2\rangle$ is easily determined, but somewhat lengthy, and we omit the explicit result here. Also of interest is the end-to-end distribution function, $P(\mathbf{R}) = \langle \delta[\mathbf{r}_N - \mathbf{R}]\rangle$, which can be obtained from Eq. (5). In order to determine the probability of an interior bond being in the "ordered" state [i.e., the fraction of residual structures, see the inset for Fig. 1(a)], we compute the interior distribution, $P_I(\mathbf{X}) = \langle \delta[\mathbf{x}-\mathbf{X}]\rangle$, so that $f_O = \int_{|\mathbf{x}|\leq c} d^3\mathbf{x}P_I(\mathbf{x})$. The radius of gyration requires a more complicated integral than the one found in Eq. (5), but we find

$$R_g^2 = \frac{Na_0^2}{6} + (\langle \mathbf{x}^2\rangle - \Delta s a_0^2)\left[\frac{\Delta s}{3N} + \frac{s_1}{N} - \left(\frac{\Delta s}{2N} + \frac{s_1}{N}\right)^2\right]. \tag{7}$$

Note that unlike the average end-to-end distance, the radius of gyration depends not only on Δs, but also on s_1.

The FRET efficiency for a system with dyes attached to $\mathbf{r}(j=0)=\mathbf{0}$ and $\mathbf{r}(i)$, $\langle E\rangle = \langle [1+(|\mathbf{r}(i)|/R_0)^6]^{-1}\rangle$, is determined from Eq. (5) as

$$E(i) = \begin{cases} E^G(i), & 0 \leq i \leq s_1 \\ \dfrac{\int_0^\infty dxdrg_1(x,r;\{s_i\})/[1+(r/R_0)^6]}{\int_0^\infty dxdrg_1(x,r;\{s_i\})}, & s_1 < i < s_2 \\ \dfrac{\int_0^\infty dxdrg_2(x,r;\{s_i\})/[1+(r/R_0)^6]}{\int_0^\infty dxdrg_2(x,r;\{s_i\})}, & s_2 \leq i \leq N, \end{cases} \tag{8}$$

where $E^G(i)$ is the FRET efficiency for a Gaussian chain with i bonds, and

$$g_1(x,r;\{s_j\}) = xr\sinh\left(\frac{3(i-s_1)xr}{\lambda a_0^2}\right)e^{-3(ix^2+\Delta s r^2)/2\lambda a_0^2 - \beta V[x]}, \tag{9}$$

$$g_2(x,r;\{s_j\}) = xr\sinh\left(\frac{3xr}{(i-\Delta s)a_0^2}\right)e^{-3x^2/2\Delta s a_0^2 - 3(x^2+r^2)/2(i-\Delta s)a_0^2 - \beta V[x]}, \tag{10}$$

$$\lambda = (s_2+s_1)i - s_1^2 - i^2. \tag{11}$$

This result allows us to compute the GSC test, after a numerical integral over r.

B. Protein L

Averages and distributions were computed using the MTM[19] which combines experimentally measured transfer free energies,[31] converged simulations and the Weighted Histogram Analysis Method (WHAM) equations.[28–30] The WHAM equations use the simulation time series of potential energy and the property of interest at various temperatures and gives a best estimate of the averages and distributions of that property. The native state ensemble (NSE) and DSE subpopulations were defined as having a structural root mean squared deviation, after least squares minimization, of less than or greater than 5 Å relative to the crystal structure for the NSE and DSE, respectively. The exact values of l_p are computed using the average R from simulations and the relationships listed in Table I.

C. Notation

Throughout the paper, exact values of all quantities are reported without superscript or subscript. For the GRM, exact values are analytically obtained or calculated by performing a one-dimensional integral numerically. For convenience, exact results for protein L refer to converged simulations. While these simulations have residual errors, the simplicity of the MTM has allowed us to calculate all properties of interest with arbitrary accuracy. The use of subscript or superscript is, unless otherwise stated, reserved for quantities that are extracted by solving Eq. (1) using the polymer models listed in Table I.

ACKNOWLEDGMENTS

We thank Sam Cho, Govardan Reddy, and David Pincus for their comments on the manuscript. E.O. thanks Guy Ziv for many useful discussions on experimental aspects of FRET measurements and analysis. This work was supported in part by grants from the NSF (No. 05-14056) to D.T., a NIH GPP Biophysics Fellowship to E.O., by the Intramural Research Program of the NIH, National Heart Lung and Blood Institute.

[1] S. E. Jackson, Folding Des. **3**, R81 (1998).
[2] A. R. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, 2nd ed. (Freeman, New York, 1999).
[3] D. K. Klimov and D. Thirumalai, J. Mol. Biol. **353**, 1171 (2005).
[4] B. Schuler and W. A. Eaton, Curr. Opin. Struct. Biol. **18**, 16 (2008).
[5] G. Haran, J. Phys.: Condens. Matter **15**, R1291 (2003).
[6] A. A. Deniz, T. A. Laurence, G. S. Beligere, M. Dahan, A. B. Martin, D. S. Chemla, P. E. Dawson, P. G. Schultz, and S. Weiss, Proc. Natl. Acad. Sci. U.S.A. **97**, 5179 (2000).
[7] A. Navon, V. Ittah, P. Landsman, H. A. Scheraga, and E. Haas, Biochemistry **40**, 105 (2001).
[8] E. Rhoades, M. Cohen, B. Schuler, and G. Haran, J. Am. Chem. Soc. **126**, 14686 (2004).
[9] K. K. Sinha and J. B. Udgaonkar, J. Mol. Biol. **353**, 704 (2005).
[10] E. V. Kuzmenkina, C. D. Heyes, and G. U. Nienhaus, Proc. Natl. Acad. Sci. U.S.A. **102**, 15471 (2005).
[11] E. Sherman and G. Haran, Proc. Natl. Acad. Sci. U.S.A. **103**, 11539 (2006).
[12] A. M. Saxena, J. B. Udgaonkar, and G. Krishnamoorthy, J. Mol. Biol. **359**, 174 (2006).
[13] A. Hoffmann, A. Kane, D. Nettels, D. E. Hertzog, P. Baumgartel, J. Lengefeld, G. Reichardt, D. A. Horsley, R. Seckler, O. Bakajin, and B. Schuler, Proc. Natl. Acad. Sci. U.S.A. **104**, 105 (2007).
[14] K. A. Merchant, R. B. Best, J. M. Louis, I. V. Gopich, and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **104**, 1528 (2007).
[15] D. Thirumalai and D. K. Klimov, Curr. Opin. Struct. Biol. **9**, 197 (1999).
[16] G. U. Nienhaus, Macromol. Biosci. **6**, 907 (2006).
[17] I. V. Gopich and A. Szabo, J. Phys. Chem. B **107**, 5058 (2003).
[18] C. Hyeon, G. Morrison, and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **105**, 9604 (2008).
[19] E. P. O'Brien, G. Ziv, G. Haran, B. R. Brooks, and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **105**, 13403 (2008).
[20] C. Magg, J. Kubelka, G. Holtermann, E. Haas, and F. X. Schmid, J. Mol. Biol. **360**, 1067 (2006).
[21] K. K. Sinha and J. B. Udgaonkar, J. Mol. Biol. **370**, 385 (2007).
[22] Q. Yi, M. L. Scalley, K. T. Simons, S. T. Gladwin, and D. Baker, Folding Des. **2**, 271 (1997).
[23] K. W. Plaxco, I. S. Millett, D. J. Segel, S. Doniach, and D. Baker, Nat. Struct. Biol. **6**, 554 (1999).
[24] D. E. Kim, C. Fisher, and D. Baker, J. Mol. Biol. **298**, 971 (2000).
[25] C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **90**, 6369 (1993).
[26] D. K. Klimov and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **97**, 2544 (2000).
[27] T. Veitshans, D. Klimov, and D. Thirumalai, Folding Des. **2**, 1 (1997).
[28] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).
[29] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992).
[30] J. Shea, Y. D. Nochomovitz, Z. Guo, and C. L. Brooks, J. Chem. Phys. **109**, 2895 (1998).
[31] M. Auton and D. W. Bolen, Biochemistry **43**, 1329 (2004).