Published in final edited form as:

J Mol Biol. 2018 August 03; 430(16): 2540–2553. doi:10.1016/j.jmb.2018.03.007.

An Extended Guinier Analysis for Intrinsically Disordered Proteins

Wenwei Zheng^a and Robert B. Best^b

^aCollege of Integrative Sciences and Arts, Arizona State University, Mesa, Arizona 85212, USA

^bLaboratory of Chemical Physics, National Institute of Diabetes Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

Abstract

Guinier analysis allows model-free determination of the radius of gyration (R_g) of a biomolecule from X-ray or neutron scattering data, in the limit of very small scattering angles. Its range of validity is well understood for globular proteins, but is known to be more restricted for unfolded or intrinsically disordered proteins (IDPs). We have used ensembles of disordered structures from molecular dynamics simulations to investigate which structural properties cause deviations from the Guinier approximation at small scattering angles. We find that the deviation from the Guinier approximation is correlated with the polymer scaling exponent ν describing the unfolded ensemble. We therefore introduce an empirical, ν -dependent, higher-order correction term, to augment the standard Guinier analysis. We test the new fitting scheme using all-atom simulation data for several IDPs and experimental data for both an IDP and a destabilized mutant of a folded protein. In all cases tested, we achieve an accuracy of the inferred R_g within $\sim 3\%$ of the true R_g . The method is straightforward to implement and extends the range of validity to a maximum qR_g of ~ 2 , versus ~ 1.1 for Guinier analysis. Compared with the Guinier or Debye approaches, our method allows data from wider angles with lower noise to be used to analyze scattering data accurately. In addition to R_g , our fitting scheme also yields estimates of the scaling exponent ν in excellent agreement with the reference ν determined from the underlying molecular ensemble.

Graphical abstract

Correspondence to: Wenwei Zheng; Robert B. Best.

Section in Appendix: Supplementary methods, figures and tables.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

Intrinsically disordered protein; SAXS; Guinier Analysis

Introduction

A number of proteins are now known to stay unfolded in physiological conditions but can still be involved in important functions like signaling or pathological pathways of diseases [1]. These so-called intrinsically disordered proteins (IDPs), and their biophysical properties, have therefore become an important research topic over the past 20 years [2, 3, 4, 5, 6]. However, the heterogeneous distribution of conformations IDPs can adopt limits the applicability of conventional structural biology methods. Small-angle X-ray scattering (SAXS) and the related small-angle neutron scattering (SANS) have emerged among the methods which are useful in characterizing disordered species [7, 8, 9, 10]. These experiments probe the distribution of interatomic distances in the ensemble, and the most fundamental parameter which can be determined from such experiments is the radius of gyration (R_g), a measure of the average degree of chain expansion. R_g has been found to be sensitive to both the solvent conditions (i.e. denaturants or salt concentrations) and sequences of IDPs (i.e. hydrophobicity and charges of the sequences) [11, 12, 13, 14], thus providing valuable structural information.

The standard method for extracting R_g from SAXS measurement starts from the SAXS intensity $I(q)$ as a function of the scattering vector q . The logarithm of the intensity can be expanded as a Taylor series in q , in which the second-order term is proportional to the mean squared radius of gyration. Truncating the high-order terms leads to the conventional Guinier approximation [15], in which R_g can be determined from the slope of a plot of $\ln[I(q)]$ versus q^2 . If the data are fitted over a range of q up to $q_{\max}$, the approximation is expected to become worse for larger $q_{\max}$. The validity of the Guinier method can be characterized by the product $q_{\max}R_g$. For folded proteins, it is known that the approximation is valid up to $q_{\max}R_g \approx 1.3$ [16, 17]. Because R_g is obtained directly from the primary data, Guinier analysis has the attractive feature of yielding a model-free estimate of R_g .

However, when applied to IDPs, the Guinier method has been found to systematically underestimate R_g for $q_{\max}R_g = 1.3$ and a much more restrictive range of at most $q_{\max}R_g \approx 1.1$ is required to obtain accurate results [18]. For $q_{\max}R_g > 1.1$, the fitted R_g becomes strongly dependent on the range of q fitted, with the degree of underestimation increasing with $q_{\max}$ [18] (note that the acceptable fitting range may also be dependent on the

properties of the protein sequence and solution conditions, as we elaborate below). Even though R_g values extracted from Guinier analysis with $q_{\max}R_g = 1.1$ are consistent with those from more sophisticated analyses of SAXS data such as ensemble fitting[18], the more restrictive $q_{\max}$ leaves only data at the smallest q , which are most affected by experimental noise. Thus, the necessary reduction in fitting range to reduce systematic error inevitably leads to much larger uncertainties in the estimated R_g [18]. This problem is exacerbated for IDPs, which often have a larger R_g relative to globular proteins, thereby further reducing the range of q from which data can be reliably analyzed by the Guinier method.

The shortcomings of Guinier analysis for studies of unfolded proteins have long been recognized [19]. One approach to overcome these difficulties has been the use of the Debye equation to analyze SAXS data of unfolded proteins[20, 19, 21]. Debye's equation is derived by combining the Gaussian chain approximation[22] and the scattering equation[23]. It is expected to work reasonably well for IDPs with a polymer scaling exponent ν close to 1/2. However scaling exponents for IDPs have been shown to span a wide range (from 1/2 to 3/5) due to both the variation of the sequence and the solvent conditions [14, 18, 24].

A more general solution to interpreting SAXS data for unfolded proteins is to model the protein at an atomistic level through generating ensembles of unfolded states with the aid of a suitable force field. This approach is also straightforwardly extended to modelling proteins with both folded and intrinsically disordered regions. These pre-generated molecular ensembles can either be reweighted or selected to fit the experimental data by a variety of schemes [25, 26, 27, 28, 18, 29], or the data can be used as a restraint in the simulation to drive the molecular ensemble to reproduce experimental averages [30, 31, 32]. These two approaches can be shown to be essentially equivalent [31, 28]. Such ensemble-type methods have been applied to all kinds of experimental data for unfolded proteins, not only SAXS, and offer a convenient way of combining different experimental data together to obtain a meaningful structural representation. Ensemble methods have become the new standard for analyzing SAXS data for IDPs [33]. In both types of method (reweighting or restraining), additional information is required for obtaining the molecular ensembles to match the SAXS data, since it has fewer degrees of freedom compared to the three-dimensional molecular structure at atomistic resolution. This is specified by the model and force field used in the simulation, which may be explicit- or implicit-solvent, all-atom or coarse-grained simulations. Naturally, the accuracy of those aspects of the ensemble not determined by the SAXS data will depend on the accuracy of the underlying force field, and these are continually being improved [34, 35, 36, 37, 38]. However, the advantage of using a force field that can reproduce sequence-dependent properties is that it can in principle capture properties of the ensemble (such as local structure formation) specific to the protein under consideration. It is thus much more suitable for describing and fitting scattering data over a wide range of scattering angles than the Guinier or Debye approximations. There is nonetheless a significant practical hurdle for practitioners, in that the method requires extensive simulations in order to be accurate, and is therefore less amenable to routine analysis.

Recently Riback et. al. [24] introduced a related, but more convenient, procedure by pre-generating molecular ensembles with different properties (specifically, the strength of

intramolecular attractions) and extracting dimensionless “molecular form factors” (MFFs). SAXS intensity can then be analyzed by finding the best MFF to describe the data. The scaling exponent ν is determined by the optimal MFF used, whereas R_g is determined from the corresponding scaling used to match the MFF to the data. The method has much in common with ensemble fitting in the sense that both start from explicitly-generated molecular ensembles, and the variation of interaction strength in generating the MFFs is analogous to the variation of temperature in implicit solvent models to generate initial solutions for reweighting [18]. The key difference is that the MFF is generated from a single sequence, but is assumed to be applicable to many different sequences; therefore new simulations do not need to be run every time, making it much more suitable for routine analysis of SAXS data. The corollary is that any sequence which deviates strongly from homopolymer structural statistics may not be well modelled by this approach. However, for the examples tested, it appeared that the results were not affected by the sequence used to generate the MFF [24].

Despite all the improvements discussed above, Guinier analysis still has many evident advantages: first it is model-free, second it is fast and convenient to perform, and third it relies only upon the small angle regime of the data set which, with the highest signal-to-noise ratio, is least influenced by background subtraction. We believe it is worth investigating where the Guinier approximation breaks down for IDPs and whether it is possible to introduce a simple correction. In the context of the Guinier framework, such a correction would be made by including higher order terms in the Taylor expansion. While higher order terms can be written in terms of the pair distance distribution, they cannot be expressed in terms of a simple property like R_g . We therefore sought to correlate empirically the magnitude of a higher order correction with structural properties of IDPs. The first property we examined was the asphericity of the protein: unlike most globular proteins which often adopt an approximately spherical shape, IDPs are aspherical, which is expected to contribute to deviations from the Guinier approximation [39]. The asphericity of an IDP can be dependent on both the sequence and solvent conditions, raising concerns when comparing R_g -s of different IDPs or IDPs in different conditions using Guinier analysis [18]. The importance of asphericity when extracting IDP structural information from Förster resonance energy transfer (FRET) or SAXS data has been stressed by several groups [40, 41, 42]. In addition to asphericity, a second factor is that IDPs adopt a distribution of conformations with different R_g s, and the more expanded and heterogeneous conformations may increase the importance of higher order terms.

The first correction term to the Guinier approximation should be $\mathcal{O}(q^4)$. We therefore first try to quantitatively characterize the impact on the fourth order term in q of variables known to be related to a breakdown of the Guinier approximation, including R_g , asphericity, variance of R_g and variance of asphericity. We find that the polymer scaling exponent ν alone determines the fourth order term sufficiently well due to its correlation with all these variables. This motivates us to introduce an empirical fourth order term dependent on ν . Similar to our recent data interpretation method for FRET [43], in which both the end-end distance and ν were determined simultaneously by introducing the correlation between the end-end distance distribution and ν using a self-avoiding walk model [44, 45, 46], we introduce the correlation between R_g and ν using the same polymer model. Together with

the empirically determined fourth order term, we have developed an extended version of Guinier analysis in which both R_g and ν are fitted at the same time and no additional free parameters are introduced. We show that the new method extends the range of validity to a maximum qR_g of ~ 2 versus ~ 1.1 for Guinier analysis, a significant advantage given the typically higher quality of the additional data available in the range $1.1 < qR_g < 2$. It is much less sensitive to the fitting range of q and more accurate for determining the R_g of IDPs than either Guinier analysis or fitting the Debye model. Most importantly, the method is only a small extension to Guinier analysis, which has been the gold standard of the field for almost 80 years.

Results

Fourth order correction to Guinier analysis is significant for IDPs

Guinier analysis is based on a series expansion of SAXS intensity, leading to an expression

$$\ln \left[\frac{I(q)}{I(0)} \right] = -\frac{1}{3}q^2 r_g^2 + \mathcal{O}(q^4), \quad (1)$$

in which q is the scattering vector, $I(q)$ the SAXS intensity and $I(0)$ the intensity when $q = 0$, and the error in the approximation should be proportional to q^4 [39]. We adopt here the convention that r_g is the radius of gyration of an individual protein conformation, while $R_g = \langle r_g^2 \rangle^{1/2}$ is the observable ensemble average. There is not much of a difference between r_g and R_g for the folded ensemble, since all folded conformations have a similar r_g . However, there is a broad distribution of r_g in the unfolded ensemble, so individual conformations may have r_g very different from the average R_g . The plot of $\ln[I(q)/I(0)]$ versus q^2 is known as the Guinier plot with slope $-r_g^2/3$ for a single conformation or $-R_g^2/3$ for the unfolded ensemble of a protein, provided that q is sufficiently small so that $\mathcal{O}(q^4)$ terms are negligible.

While the coefficient of the second order term in the expansion of $\ln[I(q)/I(0)]$ has a simple analytical relation to R_g , the coefficient of the fourth order term cannot be written in terms of such a straightforward function, although it is still a function of the pair distances. For the sake of simplicity, we make the ansatz that the fourth order term of $\ln[I(q)/I(0)]$ is of the form $-\alpha q^4 R_g^4$, where we seek to parameterize the value of the coefficient α in terms of the properties of the molecular ensemble. To expose deviations from the Guinier approximations, we rearrange Eq. 1,

$$\tilde{I} = -\frac{\ln [I(q)/I(0)]}{q^2 R_g^2} = \frac{1}{3} + \alpha q^2 R_g^2 + \mathcal{O}(q^4) \quad (2)$$

in which α is the coefficient of the fourth order term and can be used as an indicator of the error of a Guinier analysis. Note that the error $\mathcal{O}(q^4)$ in this expression in fact corresponds to a sixth order correction term due to the division of everything by $q^2 R_g^2$ in comparing to Eq. 1. A positive α indicates an overestimation of R_g if the fourth order term is neglected, whereas a negative α suggests a corresponding underestimation. We therefore use plots of $\tilde{I}$ versus $q^2 R_g^2$ to assess the magnitude of the fourth order term and range of validity for Guinier analysis.

In order to probe the dependence of α on the properties of the unfolded ensemble, we use molecular data sets from implicit solvent simulations of ACTR and R17, as well as of the homopolymer poly-Val with 100 residues. ACTR is an IDP with 79 residues[47] whereas the 116-residue R17 sequence used here is a mutant of R17 spectrin domain[48] with reduced folded state stability. These molecular ensembles were generated using temperature replica exchange Monte Carlo simulations [18], as described in the Materials and Methods section. The SAXS intensity of molecular conformation in the ensemble is calculated by using FoXS[49] with default parameters. We note that the precise choice of parameters and SAXS calculation algorithm should make little difference, since we are primarily interested in low- q data which are relatively insensitive to the detailed treatment of the solvent layer around the protein: in our recent work, we have compared several implicit and an explicit treatment of solvent in calculating SAXS intensity by using CRY SOL[50], FoXS[49] and Capriqorn [51] and found that they look almost identical within range of $q < 0.1 \text{ \AA}^{-1}$.

In Fig. 1a, we plot $\tilde{I}$ versus $q^2 R_g^2$ for selected data sets of the three proteins. Each set is a canonical ensemble taken from one of the temperatures in the replica exchange simulations; the temperatures were chosen so that the scaling exponents determined from internal distance scaling span the range known to be typical for IDPs. The separate scattering curves calculated for individual conformations indicate an increasing deviation from the Guinier approximation with larger q . Most of the $\tilde{I}$ versus $q^2 R_g^2$ curves are linear up to $q^2 R_g^2 \sim 4$, at which point the higher order terms in the expansion start to become significant. We define α to be the slope of the least square best fit line to each curve in the range of $q^2 R_g^2 < 4$ (equivalently $q R_g < 2$). For the rest of this paper we discuss a fitting scheme using a fourth order term within this range, without considering sixth order or higher terms. The slope α of these curves can be both positive or negative for individual conformations. However when calculating ensemble average curves, we always obtain a negative α , indicating an underestimation of R_g for ν spanning from 0.41 to 0.63.

There is also a systematic decrease of α with increasing ν , such that a more expanded ensemble will give a more negative α , and a greater underestimation of R_g from a Guinier analysis. In Fig. 1b, we show α as a function of ν . For three polypeptides with quite different chain length and sequence properties there is a similar relationship between α and ν . The value of α decreases from approximately -0.007 to -0.020 when varying ν from 0.41 to 0.63. The relative error of R_g due to the fourth order term can be estimated by rearranging Eq. 2 and truncating the higher order terms,

$$R_{g, \text{Guinier}}^2(q) = -\frac{3 \ln(I(q)/I(0))}{q^2} = R_g^2(1 + 3\alpha q^2 R_g^2), \quad (3)$$

This equation shows how a “single-point” Guinier estimate, $R_{g, \text{Guinier}}^2(q)$, of R_g at a particular q value deviates from the true R_g . The error of Guinier analysis is therefore proportional to both qR_g and α . The dependence on α shows explicitly that the cut-off value of qR_g to obtain a given accuracy of R_g will be dependent on α . For example, if the peptide has $\nu \approx 0.6$ and therefore α of about -0.020 in (Fig. 1b), this corresponds to a 10.1% underestimation of R_g if the data is fit to $q_{\max}R_g = 1.3$ and 7.3% for $q_{\max}R_g = 1.1$ using Eq. 3. On the other hand, for the smallest ν (0.41) of ACTR, α is only -0.007, so that fitting up to $q_{\max}R_g = 1.3$ corresponds to an error of 3.5% and fitting up to $q_{\max}R_g = 1.1$ an error of 2.5%. This is the reason why Guinier analysis is accurate for folded proteins (typically with $\nu < 0.4$) with a fitting range of $q_{\max}R_g = 1.3$, but for IDPs the error introduced by the analysis method becomes significant.

Multiple sequences help define an empirical correction term

Our results so far suggest that the coefficient α in our empirical fourth order correction term to the Guinier fit is correlated with the scaling exponent ν for several different sequences. In order to test whether α can be reasonably modeled as a function of ν alone, and that any additional dependence on chain length is small, we ran further simulations using repeated sequences of the first 40 residues of ACTR, so that the ensembles we have span a chain length from 40 to 200 amino acids, but with the same sequence composition (see details in Materials and Methods). We then perform the same analysis by obtaining α for each set of ensembles with different ν . In Fig. 2, we show α as a function of ν for all the sequences tested. There is a clear correlation between α and ν , which is best fit by the linear equation,

$$\alpha = -0.0479 \cdot (\nu - 0.212) \quad (4)$$

If we introduce a fourth order correction term combining Eq. 4 with 2, we obtain

$$\ln\left(\frac{I(q)}{I(0)}\right) = -\frac{1}{3}q^2R_g^2 + 0.0479(\nu - 0.212)q^4R_g^4. \quad (5)$$

There are three free parameters in this equation (i.e. R_g , $I(0)$ and ν) in contrast to the two parameters (R_g , $I(0)$) in the standard Guinier analysis. This would seem to make the fitting scheme less robust due to the already limited number of experimental data points in the relevant range of q for many IDPs. However, encouraged by our recent success in using a self-avoiding walk (SAW) model[44, 45] to interpret experimental FRET data from IDPs[43], we sought to introduce a dependence between R_g and ν based on the SAW,

$$R_g = \sqrt{\frac{\gamma(\gamma+1)}{2(\gamma+2\nu)(\gamma+2\nu+1)}} b N^\nu \quad (6)$$

in which $\gamma \approx 1.1615$ [52], the prefactor b has been determined to be 0.55 nm for proteins [14, 43] and N is the number of peptide bonds in the chain (i.e. one less than the number of residues). With both Eq. 5 and 6, there are only two free parameters, due to the dependence of ν on R_g . We term the self-consistent fitting of Eq. 5 and 6 to scattering data $I(q)$ “Extended Guinier Analysis”.

One source of error for using Eq. 5 would be from the deviation of linear relation between α and ν . The average relative error of the linear fit is about 10% on α . We can estimate an upper bound for the errors we have introduced using a linear relation between ν and α . If we fit to the range of $q_{\max} R_g = 2$, and assume we have $\alpha = -0.02$ (almost the maximum α for the ensembles we tested), we will get an error of 2.4% for R_g , much smaller than $\sim 10\%$ error when fitting to $q_{\max} R_g = 1.3$ using Guinier analysis. In addition, the fourth order term is able to capture the non-linear behavior between $\ln(I(q)/I(0))$ and q^2 , which is not taken into account by the second order term. This in principle makes the fitting more robust to the range of q used.

To illustrate the quality of the extended fitting scheme in comparison to the Guinier fit, we use the experimental SAXS data of R17 [18] as an example. We show in Fig. 3 the three fitting schemes, Guinier, Debye and extended Guinier analysis, on the Guinier plot. First, it is clear that the Guinier curve fits the data excellently for $q_{\max} R_g = 1.1$. However the Guinier fitting curve starts to deviate clearly from the experimental data points after $q R_g = 1.3$. Second, the Debye equation, defined by

$$\frac{I(q)}{I(0)} = \frac{2}{q^4 R_g^4} (q^2 R_g^2 - 1 - \exp[-q^2 R_g^2]) \quad (7)$$

fits the data set very well and is empirically known to fit until $q_{\max} R_g \approx 9$ [19] because there are higher order terms in the equation to match the curve. However, a good fit does not guarantee an accurate value of R_g . Third, the extended Guinier analysis fits until $q_{\max} R_g \approx 2.5$, which is larger than the range we used to obtain the slope α ($q_{\max} R_g = 2.0$). We will compare the accuracy of the three fitting schemes in the next section.

R_g is more accurately determined by extended Guinier analysis

To compare the accuracy of different fitting schemes, we first utilize a set of all-atom explicit-solvent simulation data for different IDP sequences (detailed in Table S1), which was not used in any way in parameterizing our model. We can then calculate both the R_g and SAXS intensity from the different molecular snapshots in the trajectory. The SAXS intensity can be analyzed by using different fitting schemes whereas the R_g calculated directly from the simulation can be used as a reference for comparing with the R_g obtained from the

analysis. Using synthetic data has the advantage of eliminating noise obtained in experimental data acquisition. However to mimic the typical properties of experimental data sets, where data very close to the beam stop is unusable, we always start fitting from $q = 0.01 \text{ \AA}^{-1}$.

In Fig. 4a, we show the comparison of R_g from simulation and from the analysis with three different fitting ranges including $q_{\max} R_g = 1.1, 1.3$ and 2.0 . The deviation of R_g from different methods is not immediately clear, due to the large range of R_g spanned by the different proteins. A general trend is that Debye's equation systematically gives a slightly larger R_g whereas Guinier analysis gives a slightly smaller R_g , compared with the reference. The size of the deviation also tends to increase with the range of q fitted, when using Guinier analysis. However both Debye's equation and extended Guinier analysis are relatively insensitive to the fitting range used (up to at least $q_{\max} R_g = 2$, deviations are of course expected if the fitting range is wide enough). We further show the errors of the methods using three different metrics to quantify agreement with reference R_g , including χ^2 , root-mean-square error (RMSE) and mean signed relative error (MSRE) (see Materials and Methods section for definition) in Fig. 4c. Extended Guinier analysis always performs the best, with Debye's equation being the second most accurate. Guinier analysis starts to introduce more than 3% RMSE when $q_{\max} R_g$ is larger than 1.3 and always underestimates the R_g with a negative MSRE. By contrast, Debye's equation introduces about 3% RMSE with $q_{\max} R_g = 2.0$ and always overestimates the R_g with a positive MSRE. The extended Guinier analysis introduces less than 1.1% RMSE and 0.9% MSRE for all the three fitting ranges. In addition, extended Guinier analysis simultaneously determines a value for ν which is in close agreement with that directly obtained from internal distance scalings in the trajectories.

We next benchmark the fitting scheme using experimental data sets. In a recent publication, we have obtained both FRET and SAXS data for ACTR and R17 [18]. An ensemble fitting method was used to find an ensemble of structures matching both experimental data sets. Ensemble fitting helps to avoid most of the approximations in determining R_g , and gives a more accurate estimate of R_g from multiple experimental sources. The R_g calculated from ensemble fitting therefore serves as the reference for the test, while we attempt to fit the original SAXS data with our model. In Fig. 5 we show the R_g obtained by using Guinier, Debye or extended Guinier analysis and compare those with the R_g obtained by using ensemble fitting. The corresponding χ^2 , RMSE and MSRE are in Table 1. A similar set of comparisons by using the ensembles determined from only SAXS experimental data can be found in Fig. S1 and Table S2, although results are highly similar to the ones using both FRET and SAXS experimental data.

For Guinier analysis, as we have shown in a previous publication [18], if we use a minimal fitting range, such that $q_{\max} R_g = 1.1$, it is possible to obtain a R_g close to that determined by the ensemble method (Fig. 5). With a fitting range of $q_{\max} R_g = 1.1$, the RMSE is 4.0%. For such small $q_{\max} R_g$, the errorbars of R_g determined become large (in some cases exceeding 30%) due to the limited number of data points available, making the results less useful. For the many IDPs which are longer than the two proteins used here (i.e. ACTR with 79 residues and R17 with 108 residues), and therefore a larger R_g , the number of data points within the

range of $q_{\max}R_g = 1.1$, will be even smaller, making fitting into such small $q_{\max}R_g$ range impossible. Moreover, when we increase the fitting range to $q_{\max}R_g = 1.3$ as was sometimes done for IDPs, χ^2 increases two-fold and RMSE to 5.5%. Guinier analysis always introduces a negative MSRE consistent with the previous test using all-atom simulation data, suggesting the R_g is always underestimated with the method.

Debye's equation has been shown to fit nicely to the raw SAXS intensity in Fig. 3, and we find the same here. Despite this good fit, we find that using Debye's equation introduces a positive MSRE and overestimates the R_g , especially for high denaturant concentrations when the scaling exponent is close to 0.6 (Debye's equation was derived for a Gaussian chain and would be expected to be most accurate where $\nu \approx 0.5$). Debye's equation is less sensitive to the fitting range, though with an RMSE comparable to Guinier analysis.

Our extended Guinier analysis estimates R_g with RMSE close to 3% in all the cases, performing better than both Guinier analysis and Debye's equation. The errors are quite insensitive to the fitting range. In fact, we find that for $q_{\max}R_g$ between 1.3 and 2.0, there is little difference in the R_g obtained, suggesting that the empirical fourth order term characterizes the change of the SAXS intensity due to ν reasonably well. Both the accuracy and robustness of fitting have been improved.

Discussion

We have found that the magnitude of our empirical fourth order correction term α is correlated with the polymer scaling exponent, ν , i.e. with the relative solvent quality. However, we were interested to know what molecular properties were most correlated with the magnitude of the deviation from the Guinier approximation, as captured by α . We investigated several candidate descriptors, including R_g^2 itself, the asphericity (δ^*) [53] of the protein, whose relation to the Guinier analysis has been discussed before [39], as well as the variance of $r_g^2(\sigma(r_g^2))$ and variance of the asphericity ($\sigma(\delta^*)$). We note that of course R_g^2 contributes directly to the error of the Guinier approximation, as is evident from Eq. 3; here we consider only *additional* contributions to the correction factor α itself. It is not helpful to consider the variation of α with each of these parameters by using the implicit solvent ensembles of ACTR and R17 in Fig. 1, since the average of each parameter varies systematically with the scaling exponent, as shown in Fig. S2, and therefore all appear correlated. To circumvent this problem, we have created sub-ensembles of structures such that we can keep one variable approximately fixed and vary the other, essentially by creating two-dimensional histograms of the structures using r_g^2 and δ^* . (as described in Methods and verified in Fig. S3). Thus for each bin in our histogram, we have a narrow ensemble of structures $\{X_i\}$ such that $r_g^2(X_i) \approx R_g^2$, and $\delta^*(X_i) \approx \langle \delta^* \rangle$, where R_g^2 and $\langle \delta^* \rangle$ are respectively the averages of r_g^2 and δ^* in that bin. The results, presented in Fig. 6a-b, show clearly that the asphericity $\langle \delta^* \rangle$ is the major determinant of α . There is almost no dependence of α on R_g^2 itself, apart from a small variation at low R_g^2 , which is most likely due to the limited number of structures observed at these lower R_g^2 values in the original simulations. Using our

histograms, we can also systematically study the effect of broadening the ensemble in a particular dimension, by selecting sub-ensembles in which one dimension is still fixed, while configurations are selected with a Gaussian distribution in the other dimension. The results of varying the standard deviation of these distributions are shown in Fig. 6c-d. Despite the large dependence of α on the mean δ^* seen in Fig. 6b, there is no dependence of α on $\sigma(\delta^*)$ for these broadened ensembles. This is essentially a consequence of integrating the product of an even (Gaussian) function $P(\delta^* - \langle\delta^*\rangle)$, and an odd (linear) function, $\alpha(\delta^*) - \alpha(\langle\delta^*\rangle)$. We do, however, find a dependence of α on the standard deviation of the distribution of $\sigma(r_g^2)$, i.e. for a given mean R_g^2 , a broader distribution of radius of gyration leads to a greater deviation from the Guinier approximation. This may seem slightly counterintuitive given that the initial structures were all chosen to have very similar α . Here, the nonlinear dependence of the correction term on r_g^2 means that the magnitude of the correction depends more on the configurations with larger r_g^2 . Hence the correction is more than would be expected based on the mean R_g^2 ; for structures chosen with fixed α_0 , the effective α is $\alpha = \alpha_0[1 + \sigma^2(r_g^2)/(R_g^2)^2]$. This is exactly the quadratic variation with $\sigma(r_g^2)$ that is observed in (c).

Thus there are two main factors that will cause a greater breakdown of the Guinier approximation as solvent quality is improved (e.g. by adding chemical denaturant), manifested as an underestimation of R_g [18]. First, as evident in Eq. 2, the chain expansion and larger R_g place stricter limits on the range of Guinier validity. The second contribution to the magnitude of the correction in Eq. 2 is α . As we have just shown, the increase of asphericity associated with the larger ν in a better quality solvent will also increase the magnitude of the fourth order correction term, leading to an even greater underestimation of R_g if using the Guinier approximation.

We note that an important practical consideration when fitting SAXS data is the choice of the fitting range, defined by the maximum q , $q_{\max}$ used in the fitting. Because the Guinier approximation is based on a Taylor expansion of the scattering equation in qR_g , the acceptable fitting range should be quantified by the dimensionless variable $q_{\max}R_g$ instead of $q_{\max}$, and the same is true for our extended Guinier approach. This is important because a different $q_{\max}$ will need to be used when analyzing the data from sequences with different R_g . This presents us with an obvious quandary, since before fitting we do not yet know R_g . Our solution is to scan the fitting range by stopping at different $q_{\max}$ and obtain R_g as a function of $q_{\max}$. We then calculate $q_{\max}R_g$ and take the maximum fitting range with $q_{\max}R_g$ smaller than the criterion we set (Eg. $q_{\max}R_g=1.1, 1.3$ or 2.0 tested). For all the investigations shown in this study, we have followed this procedure, assuming we do not know the reference R_g beforehand. In practice when applying the extended Guinier analysis to IDP data sets, we suggest the same procedure of scanning the fitting range and stopping at the largest q with $q_{\max}R_g < 2$. The additional cost of scanning is negligible compared with ensemble fitting methods due to the ease of fitting an empirical equation.

The importance of accurate data analysis was highlighted by the recent controversy between the interpretation of FRET and SAXS experiments on unfolded proteins – FRET

experiments had shown proteins to expand in chemical denaturant [54, 55, 56, 57], while several SAXS studies had found no detectable expansion [58, 59, 60]. We, and others, have recently addressed the causes of this controversy [61, 40, 41, 18, 42, 24]. Part of the resolution lies in the fact that earlier SAXS experiments were performed on stable proteins whose unfolded states could only be studied at high denaturant concentrations, while most of the expansion with denaturant occurs at low denaturant concentrations [18, 42, 24]. In addition, at high denaturant concentrations, the end-end distance monitored by FRET continues to increase, while the R_g monitored by SAXS varies much less – a simple consequence of the polymeric properties of the chain. However, there was also a contribution from analysis methods: use of a Gaussian chain model for FRET tended to overestimate the R_g at high denaturant concentrations [61, 40, 18], whereas Guinier analysis for SAXS tends to underestimate the R_g changes for IDPs at high denaturant concentrations [18].

While ensemble fitting [28] is a powerful and general alternative which overcame these problems [18], it is more complicated to implement and comes with a significant computational cost. Simpler, but still accurate, methods are therefore desirable. In a recent publication [43], we presented and tested an improved method of FRET analysis based on the distance distribution of a self-avoiding walk model [46]. In this work, we employ the same model in conjunction with an empirical higher order correction to the Guinier approximation, to generate a simple to implement, yet accurate method for determining the radius of gyration of disordered proteins from scattering data. Whilst we have demonstrated the superiority of our method for the analysis of unfolded and disordered proteins, it is nonetheless a solution specific to disordered proteins. For proteins containing a more complex mixture of folded and disordered regions, it is still necessary to use either Guinier or – preferably – ensemble fitting for analysis.

Materials and Methods

Ensemble preparation

The ensembles of ACTR and R17 covering a variety of ν are the same as the ones used in our previous work for fitting the experimental data [18]. Therefore we briefly discuss the way of obtaining these ensembles here. The implicit-solvent ABSINTH [34] model and the program CAMPARI are used with the OPLS model and a temperature-independent implicit solvent. Temperature replica exchange Monte Carlo was run for 1.3 million steps, spanning a range of 280 to 2446 K with 64 replicas, starting from a fully extended configuration. At each temperature, the last 1.0 million steps were evenly recorded as 10,000 protein configurations. As shown in Figure S4 of the original publication [18], ν spans a range of 0.41 to 0.63 with 64 replicas, providing a good database of conformations for creating ensemble with unique features in this paper.

With the same simulation protocol, we have also generated a series of ensembles for the repeated sequences of the first 40 residues of ACTR sequence, so that we have an artificial peptide of $[\text{ACTR40}]_n$ in which n equals 1, 2, 3, 4 and 5. The same number of replicas and temperature setup are used and the same number of configurations are collected for each temperature as compared to ACTR and R17.

Error analysis

We introduce three different variables to compare different analysis methods. The first is χ^2 defined as $\chi^2 = \langle (R_{g,\text{analysis}} - R_{g,\text{reference}})^2 / (\sigma_{\text{reference}}^2) \rangle$. Here the reference values can be the values calculated directly from the simulations if we use synthetic data (i.e. all-atom simulation data), or the values obtained from ensembles fitted to the experimental measurements if we use experimental data for the test. The second is the root-mean-square error (RMSE), defined as $\text{RMSE} = (\langle (R_{g,\text{analysis}} - R_{g,\text{exp}})^2 / (R_{g,\text{exp}})^2 \rangle)^{1/2}$. The value is the root mean square relative error and is shown in percentage. The third is the mean signed relative error (MSRE) defined as $\text{MSRE} = \langle (R_{g,\text{analysis}} - R_{g,\text{exp}}) / R_{g,\text{exp}} \rangle$. Value of RMSE and MSRE are given as percentages.

Definition of asphericity

Asphericity δ^* is defined here as [53]

$$\delta^* = 1 - 3 \frac{(\lambda_1 \lambda_2 + \lambda_2 \lambda_3 + \lambda_1 \lambda_3)}{(\lambda_1 + \lambda_2 + \lambda_3)^2} \quad (8)$$

where $\{\lambda_i\}$ are the eigenvalues of the gyration tensor, defined for a given conformation $\mathbf{R} = r_1, r_2, \dots, r_N$ with center of geometry $\mathbf{r}_{\text{CG}}$ as

$$\mathcal{T} = \frac{1}{N} \sum_{i=1}^N (\mathbf{r}_i - \mathbf{r}_{\text{CG}}) \otimes (\mathbf{r}_i - \mathbf{r}_{\text{CG}}). \quad (9)$$

Sub-ensemble selection

Starting from the initial ensembles of ACTR described above, we sought to create four sets of sub-ensembles by selecting configurations in which only one of the four variables (R_g , δ^* , $\sigma(r_g)$, and $\sigma(\delta^*)$) vary. Before selecting the configurations, we always combine all the configurations from different temperatures together as a configuration database. For varying R_g , we first make a histogram on δ^* with 20 bins. For each δ^* bin we make a second histogram on R_g with 10 bins. This creates a set of sub-ensembles with fixed δ^* and varying R_g . Within each sub-ensemble, the r_g and δ^* are very similar to each other and therefore the variance of each value is extremely small (Fig. S3 leftmost). For varying δ^* , a similar idea is applied except that we first make a histogram (20 bins) on R_g and then a second histogram (10 bins) on δ^* (Fig. S3 left middle).

To vary $\sigma(r_g)$, we first make a histogram on δ^* with 20 bins and select the bin with the largest number of configurations. Now we want to create different ensembles with the same $R_g^2 = \langle r_g^2 \rangle$, but different $\sigma(r_g^2)$ using these configurations with fixed δ^* . We make a second histogram on R_g with 15 bins using these configurations. For a pre-calculated Gaussian distribution of r_g with a defined variance, we can calculate the number of configurations we

need for each of these 15 bins. For one (the most populated) of the 15 bins in r_g^2 , all the configurations are used, whereas for the remaining 14 bins, we randomly choose the number of structures required to give a Gaussian distribution in r_g^2 . We then repeat the process several times until we have 10 new ensembles with different $\sigma(r_g)$ but the same R_g (Fig. S3 right middle). For varying $\sigma(\delta^*)$, we follow the same procedure as used for creating ensembles with varying $\sigma(r_g)$ except that we first make a histogram (20 bins) on r_g and then the second histogram (15 bins) on δ^* .

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Attila Szabo for pressing us to consider this problem and for many helpful suggestions. We also thank Joshua Riback and Tobin Sosnick for comments on the manuscript. WZ and RBB were supported by the Intramural Research Program of the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health. WZ thanks Arizona State University for the start-up support. This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>).

References

1. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol.* 2015; 16(1):18–29. [PubMed: 25531225]
2. Sigler PB. Acid blobs and negative noodles. *Nature.* 1988; 333(6170):210–212. [PubMed: 3367995]
3. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol.* 1999; 293(2):321–331. [PubMed: 10550212]
4. Uversky VN, Gillespie JR, Fink AL. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins.* 2000; 41:415–427. [PubMed: 11025552]
5. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, et al. Intrinsically disordered protein. *J Mol Graph Model.* 2001; 19(1):26–59. [PubMed: 11381529]
6. Tompa P. Intrinsically unstructured proteins. *Trends Biochem Sci.* 2002; 27:527–533. [PubMed: 12368089]
7. Lipfert J, Doniach S. Small-angle x-ray scattering from rna, proteins, and protein complexes. *Annu Rev Biophys Biomol Struct.* 2007; 36:307–327. [PubMed: 17284163]
8. Bernadó P, Svergun DI. Structural analysis of intrinsically disordered proteins by small-angle x-ray scattering. *Mol Biosyst.* 2012; 8(1):151–167. [PubMed: 21947276]
9. Martin EW, Holehouse AS, Grace CR, Hughes AJ, Pappu RV. Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation. *J Am Chem Soc.* 2016; 138:15323–15335. [PubMed: 27807972]
10. Cragnell C, Durand D, Cabane B, Skepö M. Coarse-grained modeling of the intrinsically disordered protein histatin 5 in solution: monte carlo simulations in combination with saxs. *Proteins.* 2016; 84:777–791. [PubMed: 26914439]
11. Hoffmann A, Kane A, Nettels D, Hertzog DE, Baumgärtel P, Lengefeld J, Reichardt G, Horsley DA, Seckler R, Bakajin O, et al. Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proc Natl Acad Sci USA.* 2007; 104(1):105–110. [PubMed: 17185422]
12. Mao AH, Crick SL, Vitalis A, Chicoine C, Pappu RV. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc Natl Acad Sci USA.* 2010; 107:8183–8188. [PubMed: 20404210]

13. Das RK, Pappu RV. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci USA*. 2013; 110:13392–13397. [PubMed: 23901099]
14. Hofmann H, Soranno A, Borgia A, Gast K, Nettels D, Schuler B. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc Natl Acad Sci USA*. 2012; 109:16155–16160. [PubMed: 22984159]
15. Guinier A. La diffraction des rayons x aux tres petits angles: applications a l'etude de phenomenes ultramicroscopiques. *Ann Phys*. 1939; 12:161–237.
16. Putnam CD, Hammel M, Hura GL, Tainer JA. X-ray solution scattering (saxs) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q Rev Biophys*. 2007; 40(3):191–285. [PubMed: 18078545]
17. Kikhney AG, Svergun DI. A practical guide to small angle x-ray scattering (saxs) of flexible and intrinsically disordered proteins. *FEBS Lett*. 2015; 589:2570–2577. [PubMed: 26320411]
18. Borgia A, Zheng W, Buholzer K, Borgia MB, Schuler A, Hofmann H, Soranno A, Nettels D, Gast K, Grishaev A, Best RB, Schuler B. Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. *J Am Chem Soc*. 2016; 138(36):11714–11726. [PubMed: 27583570]
19. Calmettes P, Durand D, Desmadril M, Minard P, Receveur V, Smith J. How random is a highly denatured protein? *Biophys Chem*. 1994; 53(1-2):105–113. [PubMed: 17020841]
20. Debye P. Molecular-weight determination by light scattering. *J Phys Chem*. 1947; 51(1):18–32.
21. Sim AY, Lipfert J, Herschlag D, Doniach S. Salt dependence of the radius of gyration and flexibility of single-stranded dna in solution probed by small-angle x-ray scattering. *Phys Rev E*. 2012; 86(2):021901.
22. Flory PJ. The configuration of real polymer chains. *J Chem Phys*. 1949; 17:303–310.
23. Debye P. Zerstreuung von röntgenstrahlen. *Ann Phys*. 1915; 351(6):809–823.
24. Riback JA, Bowman MA, Zmyslowski AM, Knoverek CR, Jumper JM, Hinshaw JR, Kaye EB, Freed KF, Clark PL, Sosnick TR. Innovative scattering analysis shows that hydrophobic proteins are expanded in water. *Science*. 2017; 358:238–241. [PubMed: 29026044]
25. Lange OF, Lakomek NA, Farés C, Schröder GF, Walter KE, Becker S, Meiler J, Grubmüller H, Griesinger C, de Groot BL. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*. 2008; 320:1471–1475. [PubMed: 18556554]
26. Boura E, Yicki BR, Herrick DZ, Chung HS, Vecer J, Eaton WA, Cafiso DS, Hummer G, Hurley JH. Solution structure of the escrt-i complex by small-angle x-ray scattering, epr, and fret spectroscopy. *Proc Natl Acad Sci USA*. 2011; 108(23):9437–9442. [PubMed: 21596998]
27. Schwalbe M, Ozenne V, Bibow S, Jaremko M, Jaremko L, Gajda M, Jensen MR, Biernat J, Becker S, Mandelkow E, et al. Predictive atomic resolution descriptions of intrinsically disordered htau40 and α -synuclein in solution from nmr and small angle scattering. *Structure*. 2014; 22(2):238–249. [PubMed: 24361273]
28. Hummer G, Köfinger J. Bayesian ensemble refinement by replica simulations and reweighting. *J Chem Phys*. 2015; 143(24):243150. [PubMed: 26723635]
29. Brookes DH, Head-Gordon T. Experimental inferential structure determination of ensembles for intrinsically disordered proteins. *J Am Chem Soc*. 2016; 138(13):4530–4538. [PubMed: 26967199]
30. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M. Simultaneous determination of protein structure and dynamics. *Nature*. 2005; 433:128–132. [PubMed: 15650731]
31. Roux B, Weare J. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J Chem Phys*. 2013; 138:084107. [PubMed: 23464140]
32. Allison JR, Rivers RC, Christodoulou JC, Vendruscolo M, Dobson CM. A relationship between the transient structure in the monomeric state and the aggregation propensities of α -synuclein and β -synuclein. *Biochemistry*. 2014; 53(46):7170–7183. [PubMed: 25389903]
33. Hub JS. Interpreting solution x-ray scattering data using molecular simulations. *Curr Opin Struct Biol*. 2018; 49:18–26. [PubMed: 29172147]

34. Vitalis A, Pappu RV. ABSINTH: A new continuum solvent model for simulations of polypeptides in aqueous solutions. *J Comput Chem*. 2008; 30:673–699.
35. Ozenne V, Bauer F, Salmon L, Huang Jr, Jensen MR, Segard S, Bernadó P, Charavay C, Blackledge M. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*. 2012; 28(11):1463–1470. [PubMed: 22613562]
36. Best RB, Zheng W, Mittal J. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J Chem Theor Comput*. 2014; 10:5113–5124.
37. Piana S, Donchev AG, Robustelli P, Shaw DE. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J Phys Chem B*. 2015; 119:5113–5123. [PubMed: 25764013]
38. Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, Grubmüller H, MacKerell A. Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods*. 2017; 14(1):71–73. [PubMed: 27819658]
39. Feigin LA, Svergun DI. Structure analysis by small-angle X-ray and neutron scattering. Springer; New York: 1987.
40. Song J, Gomes GN, Gradinaru CC, Chan HS. An adequate account of excluded volume is necessary to infer compactness and asphericity of disordered proteins by forster resonance energy transfer. *J Phys Chem B*. 2015; 119(49):15191–15202. [PubMed: 26566073]
41. Song J, Gomes GN, Shi T, Gradinaru CC, Chan HS. Conformational heterogeneity and fret data interpretation for dimensions of unfolded proteins. *Biophys J*. 2017; 113:1012–1024. [PubMed: 28877485]
42. Fuertes G, Banterle N, Ruff KM, Chowdhury A, Mercadante D, Koehler C, Kachala M, Girona GE, Milles S, Mishra A, et al. Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in saxs vs. fret measurements. *Proc Natl Acad Sci USA*. 2017; 114:E6342–E6351. [PubMed: 28716919]
43. Zheng W, Zerze G, Borgia A, Mittal J, Schuler B, Best R. Inferring properties of disordered chains from fret transfer efficiencies. *J Chem Phys*. in press.
44. Fisher ME. Shape of a self-avoiding walk or polymer chain. *J Chem Phys*. 1966; 44(2):616–622.
45. Des Cloizeaux J. Lagrangian theory for a self-avoiding random chain. *Phys Rev A*. 1974; 10(5): 1665.
46. de Gennes PG. Scaling Concepts in Polymer Physics. Cornell University Press; 1978.
47. Demarest SJ, Martinez-Yamout M, Chung J, Chen H, Xu W, Dyson HJ, Evans RM, Wright PE. Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature*. 2002; 415:549–553. [PubMed: 11823864]
48. Borgia A, Wensley BG, Soranno A, Nettels D, Borgia MB, Hoffmann A, Pfeil SH, Lipman EA, Clarke J, Schuler B. Localizing internal friction along the reaction coordinate of protein folding by combining ensemble and single-molecule fluorescence spectroscopy. *Nat Commun*. 2012; 3:1195. [PubMed: 23149740]
49. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. Accurate saxs profile computation and its assessment by contrast variation experiments. *Biophys J*. 2013; 105(4):962–974. [PubMed: 23972848]
50. Svergun D, Barberato C, Koch MHJ. CRY SOL - a program to evaluate X-ray solution scattering of biological molecules from atomic coordinates. *J Appl Cryst*. 1995; 28:768–773.
51. Köfinger J, Hummer G. Atomic-resolution structural information from scattering experiments on macromolecules in solution. *Phys Rev E*. 2013; 87(5):052712.
52. Le Guillou J, Zinn-Justin J. Critical exponents for the n-vector model in three dimensions from field theory. *Phys Rev Lett*. 1977; 39(2):95.
53. Steinhauser MO. A molecular dynamics study on universal properties of polymer chains in different solvent qualities. part i. a review of linear chain properties. *J Chem Phys*. 2005; 122(9): 094901. [PubMed: 15836175]
54. Schuler B, Lipman EA, Eaton WA. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*. 2002; 419:743–747. [PubMed: 12384704]

55. Sherman E, Haran G. Coil-globule transition in the denatured state of a small protein. *Proc Natl Acad Sci USA*. 2006; 103:11539–11543. [PubMed: 16857738]
56. Merchant KA, Best RB, Louis JM, Gopich IV, Eaton WA. Characterizing the unfolded states of proteins using single molecule FRET spectroscopy and molecular simulations. *Proc Natl Acad Sci USA*. 2007; 104:1528–1533. [PubMed: 17251351]
57. Ziv G, Haran G. Protein folding, protein collapse, and Tanford's transfer model: lessons from single-molecule FRET. *J Am Chem Soc*. 2009; 131:2942–2947. [PubMed: 19239269]
58. Plaxco KW, Millett IS, Segel DJ, Doniach S, Baker D. Chain collapse can occur concomitantly with the rate-limiting step in protein folding. *Nature Struct Biol*. 1999; 6:554–556. [PubMed: 10360359]
59. Jacob J, Dothager RS, Thiyagarajan P, Sosnick TR. Fully reduced ribonuclease A does not expand at high denaturant concentration or temperature. *J Mol Biol*. 2007; 367:609–615. [PubMed: 17292402]
60. Yoo TY, Meisburger SP, Hinshaw J, Pollack L, Haran G, Sosnick TR, Plaxco K. Small-angle X-ray scattering and single-molecule FRET spectroscopy produce highly divergent views of the low-denaturant unfolded state. *J Mol Biol*. 2012; 418:226–236. [PubMed: 22306460]
61. O'Brien EP, Morrison G, Brooks BR, Thirumalai D. How accurate are polymer models in the analysis of Förster resonance energy transfer experiments on proteins? *J Chem Phys*. 2012; 130:124903.

Research Highlights

- Guinier analysis is often impractical for obtaining a radius of gyration for intrinsically disordered proteins due to its narrow range of validity
- The correction to Guinier analysis is significant for IDPs due to their larger radius of gyration and increased asphericity
- We have developed a simple empirical correction term parameterized on simulation data and validated it against other simulations and against experiment
- The correction is as simple to use as conventional Guinier analysis, but should allow radii of gyration to be more accurately determined

Figure 1.

Investigation of fourth order correction term to the Guinier approximation. a) Magnitude of the correction term to Guinier analysis, as a function of $q^2 R_g^2$. The blue lines show the curve for individual configurations in the ensemble whereas red lines show the ensemble averaged values. Dashed lines show the best fit linear curve with slope α (fitting between $0 < q^2 R_g^2 < 4$). Legends show the polymer scaling exponent ν determined from internal distance scaling with sequence separation for each ensemble. b) The fitted slope α of curves in (a) plotted as a function of ν . Left column shows R17, middle column ACTR and right column poly-Val.

Figure 2.

Parameterizing α in terms of ν . We plot the relation between α and ν for implicit solvent simulations with the ABSINTH model[34] for a variety of sequences (see legend). Note that each data point corresponds to a single temperature from the replica exchange simulations of each protein.

Figure 3.

Fit of different models to *low-q* experimental data. The fit of SAXS data for R17 (black) obtained in 0.58M GdmCl using different approximations, including Guinier analysis (blue, with $q_{\max}R_g = 1.1$), Debye's equation (red, with $q_{\max}R_g = 2.0$), and extended Guinier analysis (cyan, with $q_{\max}R_g = 2.0$). Vertical black lines indicate different values of qR_g , using the R_g determined by ensemble fitting, as labelled on top axis.

Figure 4.

Testing using all-atom MD simulations with explicit solvent. The R_g -s determined from Guinier analysis (blue), Debye's equation (red), and extended Guinier analysis (cyan) are compared with reference values obtained directly from the simulation trajectories. The three columns show the results for fitting ranges $q_{\max} R_g = 1.1$, 1.3 and 2.0, respectively. In (a) we compare R_g from analysis to the true simulation R_g , and in (b) the scaling exponent ν from analysis with that calculated from the internal distance scaling in the simulations. In (c), we summarize the deviation of inferred R_g from the true value using χ^2 , RMSE and MSRE.

Figure 5.

Testing on experimental data for different proteins and denaturation conditions. We show a comparison between the R_g extracted from (a) Guinier analysis, (b), Debye's equation and (c) extended Guinier analysis on different experimental data sets including R17 and ACTR in different denaturant conditions. The corresponding scaling exponent, ν , from extended Guinier analysis is shown in (d). The reference values (grey shaded regions) are from ensembles fitted to both FRET and SAXS experimental data from Ref. [18]. e) Errors estimated using χ^2 , RMSE and MSRE for using Guinier analysis with $q_{\max}R_g = 1.3$ (blue), Debye's equation with $q_{\max}R_g = 2.0$ (red) and extended Guinier analysis with $q_{\max}R_g = 2.0$ (cyan). Color code is shown in the legend: $q_{\max}R_g$ used for fitting in (a), (b), (c) and (d); or the method used for interpreting data in (e).

Figure 6.

Correlations between correction coefficient α and structural descriptors. Descriptors include mean squared radius of gyration R_g^2 (a), mean asphericity $\langle \delta^* \rangle$ (b), variance of squared radius of gyration $\sigma(r_g^2)$ (c) or variance of asphericity $\sigma(\delta^*)$ (d) for a series of ensembles selected by fixing the other three variables. Legends show the corresponding fixed R_g^2 and $\langle \delta^* \rangle$ for the ensembles used.

Table 1**Errors for FRET and SAXS**

Error	Method	$q_{\max}R_g$	RI7 GdmCl	RI7 Urea	ACTR GdmCl	ACTR Urea	Average
χ^2	Guinier	1.1	5.03	4.93	3.61	6.55	5.03
		1.3	7.46	8.08	9.56	15.75	10.21
	Debye	1.3	6.49	8.96	25.19	8.27	12.23
	Extended	2.0	7.72	17.41	20.34	10.32	13.95
		1.3	1.16	1.89	6.64	1.06	2.69
		2.0	0.67	5.10	2.13	5.29	3.30
RMSE(%)	Guinier	1.1	4.4	5.4	2.6	3.4	4.0
		1.3	5.2	5.9	4.4	6.5	5.5
	Debye	1.3	4.9	5.3	6.6	4.8	5.4
	Extended	2.0	5.4	6.7	5.3	4.5	5.5
		1.3	2.2	2.7	3.5	1.8	2.6
		2.0	1.6	4.5	2.3	4.1	3.1
MSRE(%)	Guinier	1.1	-2.9	-3.6	0.6	-2.1	-2.0
		1.3	-5.0	-5.3	-4.0	-6.1	-5.1
	Debye	1.3	4.5	4.6	6.4	4.4	5.0
	Extended	2.0	5.0	4.8	4.7	3.5	4.5
		1.3	1.4	1.1	3.0	1.0	1.6
		2.0	-0.9	-2.1	-1.6	-3.1	-1.9