Biophysical Journal Volume 106 June 2014 2474–2482

Probing the Average Local Structure of Biomolecules Using Small-Angle Scattering and Scaling Laws

Max C. Watson* and Joseph E. Curtis*
NIST Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, Maryland

ABSTRACT Small-angle neutron and x-ray scattering have become invaluable tools for probing the nanostructure of molecules in solution. It was recently shown that the definite integral of the scattering profile exhibits a scaling (power-law) behavior with respect to molecular mass. We derive the origin of this relationship, and discuss how the integrated scattering profile can be used to identify differing levels of disorder over local $\lesssim$30 Å length scales. We apply our analysis to globular and intrinsically disordered proteins.

INTRODUCTION

Proteins and polymers in solution have many common features. Like a polymer chain, proteins are composed of a long chain of monomer units. The degree of folding for both polymers and proteins is strongly influenced by their interactions with the solvent. The global structure of many proteins is consistent with Flory's scaling law for polymers (1),

$$R_g = R_0 N^{\nu}, \tag{1}$$

where R_g is the radius of gyration, N is the number of chain segments, and R_0 is the length of each segment. The value of ν lies in the range $1/3 \leq \nu \lesssim 3/5$, and depends on the nature of the polymer chain. Experimental measurements (2,3) have found that unfolded proteins exhibit the behavior of a self-avoiding random walk ($\nu = 0.588$). When interactions between the solvent and chain are sufficiently unfavorable, polymers collapse into a compact shape ($\nu = 1/3$). This prediction is in good agreement with measurements of globular proteins (4,5).

Unlike polymers, however, globular proteins adopt conformations specific to their amino-acid sequence and physiological role. Despite the overall scaling trend described by Eq. 1, the spread in the experimental data is large (see Fig. 6). Two proteins with the same number of residues can have radii of gyration that differ by an order of magnitude. Although Flory's scaling law offers a qualitative model of protein size, it does not provide a structural description that is closely obeyed by all proteins.

In this article, we demonstrate how the nanostructure of biomolecules can be accurately characterized by integrating over their small-angle scattering profile. Our analysis was applied to both globular and intrinsically disordered proteins. For nearly all cases, the integrated scattering profile scales with the number of residues in a manner similar to Eq. 1, but exhibits much less dispersion, reflecting similarities in density on length scales below $\approx$30 Å. Our work is inspired by a recent article by Rambo and Tainer (6), who discovered a scaling relationship between the mass of biomolecules and a quantity related to their integrated scattering profile. However, the investigation was largely empirical and an origin of the scaling behavior was not given. Furthermore, their interpretation relied on quantities that are undefined for disordered molecules, which lack a well-defined shape.

In addition to analyzing a wider class of proteins than Rambo and Tainer (6), we elucidate the physical meaning of the integrated scattering profile using an approach that can be applied to both compact and disordered molecules. We show that integrating the small-angle scattering profile up to a maximum wavenumber q_{m} corresponds to scanning the entire particle with a probe of radius $2\pi/q_{\mathrm{m}}$, providing structural information that cannot be directly obtained from the scattering profile itself. An approximate scaling relationship between the integrated scattering profile and number of residues for globular molecules is also derived, which is in good agreement with experimental measurements. In addition, we discuss how an individual molecule's deviation from this scaling trend can be used to quantify its degree of disorder.

THEORY

For small-angle neutron and x-ray scattering measurements, the scattering profile $I(q)$ of a molecule in a dilute solution may be written as

$$I(q) = 4\pi \int_0^D p(r)\mathrm{sinc}(qr)\mathrm{d}r, \tag{2}$$

where $\mathrm{sinc}(x) \equiv \sin(x)/x$. The magnitude q of the scattering vector is given by $q = 4\pi\sin(\theta)/\lambda$, where 2θ is the scattering angle and λ is the wavelength of the incident radiation. The value $p(r)$ is the particle's pair distribution function (7,8), which gives an effective histogram for atom pairs separated

Submitted November 6, 2013, and accepted for publication March 25, 2014.
*Correspondence: max.watson@nist.gov or joseph.curtis@nist.gov
Editor: Lois Pollack.

0006-3495/14/06/2474/9 $2.00
http://dx.doi.org/10.1016/j.bpj.2014.03.050

by a distance r. Because the atomic separation cannot be greater than the maximum dimension of the particle D, we have $p(r) = 0$ when $r > D$.

To characterize local structure, we define

$$V_c(q_m) \equiv \frac{I(0)}{\int_0^{q_m} qI(q)dq}. \quad (3)$$

The theoretical properties of V_c have been described in the case where the integral's upper limit extends to infinity (7). Because Eq. 3 contains the normalized profile $I(q)/I(0)$, $V_c(q_m)$ can be obtained when absolute intensity measurements are not available (6). For a homogeneous, rigid particle and an infinite upper limit of integration, V_c is proportional to the particle volume divided by its average chord (correlation) length (7). Rambo and Tainer (6) used these assumptions to interpret experimental measurements of V_c, although their actual integration extended to $q_m = 0.5$ Å^{-1}.

In this section, we explain the physical significance of $V_c(q_m)$ without assuming anything about the nature of the particle or the value of q_m. It will be shown that $2\pi/q_m$ describes an effective probe size for scanning the particle. $V_c(q_m \rightarrow \infty)$, therefore qualitatively differs from V_c[0.5 Å^{-1}]. We will also demonstrate how $V_c(q_m)$ provides a unique parameter for quantifying disorder and molecular shape over the length scale of the probe size. Finally, we derive an approximate scaling relationship between the integrated scattering profile and the number of residues for globular molecules.

In the general case, q_m may be understood as follows. Substituting Eq. 2 into Eq. 3 and using the relations $1 - \cos(x) = 2\sin^2(x/2)$ and $2\sin^2(q_m r/2)/r^2 = (q_m{}^2/2)\mathrm{sinc}^2(q_m r/2)$, we have

$$V_c(q_m) = \frac{2I(0)}{q_m^2 J(q_m)}, \quad (4a)$$

where

$$J(q_m) \equiv 4\pi \int_0^D p(r)\mathrm{sinc}^2\left(\frac{q_m r}{2}\right)dr. \quad (4b)$$

Expressed in this form, $J(q_m)$ and $I(q)$ closely resemble each other. The function $\mathrm{sinc}(x)$ is present in both $I(q)$ and $J(q_m)$. But in contrast to $\mathrm{sinc}(x)$ (Eq. 2), the properties of $\mathrm{sinc}^2(x)$ (Eq. 4b) make $J(q_m)$ better suited for measuring local structure (see Fig. 1). Because $\mathrm{sinc}^2(q_m r/2)$ is always positive and rapidly decays for atomic separations $r > 2\pi/q_m$, $J(q_m)$ corresponds to a sum over all atom pairs whose separation is less than $2\pi/q_m$. This can be roughly understood as $J(q_m) \approx \int_0^{2\pi/q_m} p(r)dr$ (this is approximate because $\mathrm{sinc}^2(x)$ is not a step function). The length scale $2\pi/q_m$ therefore describes an effective probe size. When the probe is much larger than the molecule ($q_m D \ll 1$), all atom pairs are counted, so that $J(q_m) \propto N^2$. When the probe is much smaller than the molecule's geometric features, only atom pairs that fit within the probe are counted, and $J(q_m) \propto N$. Because $I(0) \propto N^2$, we have $V_c = 2/q_m{}^2 \propto N^0$ and $V_c \propto N$ in these two regimes, respectively.

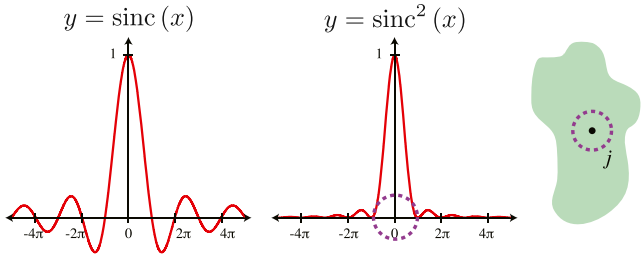

FIGURE 1 The functions $\mathrm{sinc}(x)$ and $\mathrm{sinc}^2(x)$ that appear in Eqs. 2 and 4b, respectively. In contrast to $\mathrm{sinc}(x)$, $\mathrm{sinc}^2(x)$ is always positive and decays more rapidly outside of the main envelope $|x| < \pi$. In Eq. 4b, this range corresponds to atomic separation distances less than $2\pi/q_m$. The value of q_m therefore describes the effective size of a probe (*dotted circle*) that scans over the entire molecule (*green*). Equation 4b approximately corresponds to placing the center of the probe at each atom j (*black*) and counting the number of atoms located inside the probe. To see this figure in color, go online.

The relationship between $V_c(q_m)$ and N therefore does not obey a simple scaling law. In other words, the slope on a logarithmic plot, $\partial\ln(V_c)/\partial\ln(N)$, is not constant over all values of N. Fitting data to a scaling law

$$V_c(q_m) = \alpha N^{\mu} \quad (5)$$

over a finite range in N would yield an apparent scaling exponent μ, which is equal to the average logarithmic slope over that interval: $\mu \approx \partial\ln(V_c)/\partial\ln(N)$. Although Eq. 5 is not strictly valid over all N, we will show that it is useful for describing the structure of biological molecules.

For a fixed number of atoms and a value of q_m between the $J(q_m) \propto N$ and $J(q_m) \propto N^2$ regimes, $J(q_m)$ is generally larger for compact, spherical molecules. Structures that are disordered and/or nonspherical do not contain as many atom pairs separated by a distance less than the probe size, and thus have a smaller $J(q_m)$. Because $J(q_m)$ is in the denominator of Eq. 4a, $V_c(q_m)$ increases with the level of disorder/asphericity. In Fig. 2, we visually demonstrate this property by comparing a sphere and a random coil ($\nu = 1/2$).

In principle, $J(q_m)$ may be obtained from scattering measurements based on the inferred pair distribution function $p(r)$ (9). The function $p(r)$ is calculated by taking the indirect Fourier transform (10) of $I(q)$, while a value of D must be assumed before the transformation. According to Svergun and Koch (11), D can be determined by iteratively transforming between $I(q)$ and $p(r)$. However, repeated transformations between real space and Fourier space can sometimes result in numerical artifacts. Unlike $p(r)$, $V_c(q_m)$ can be directly obtained from the scattering profile without resorting to indirect Fourier transforms.

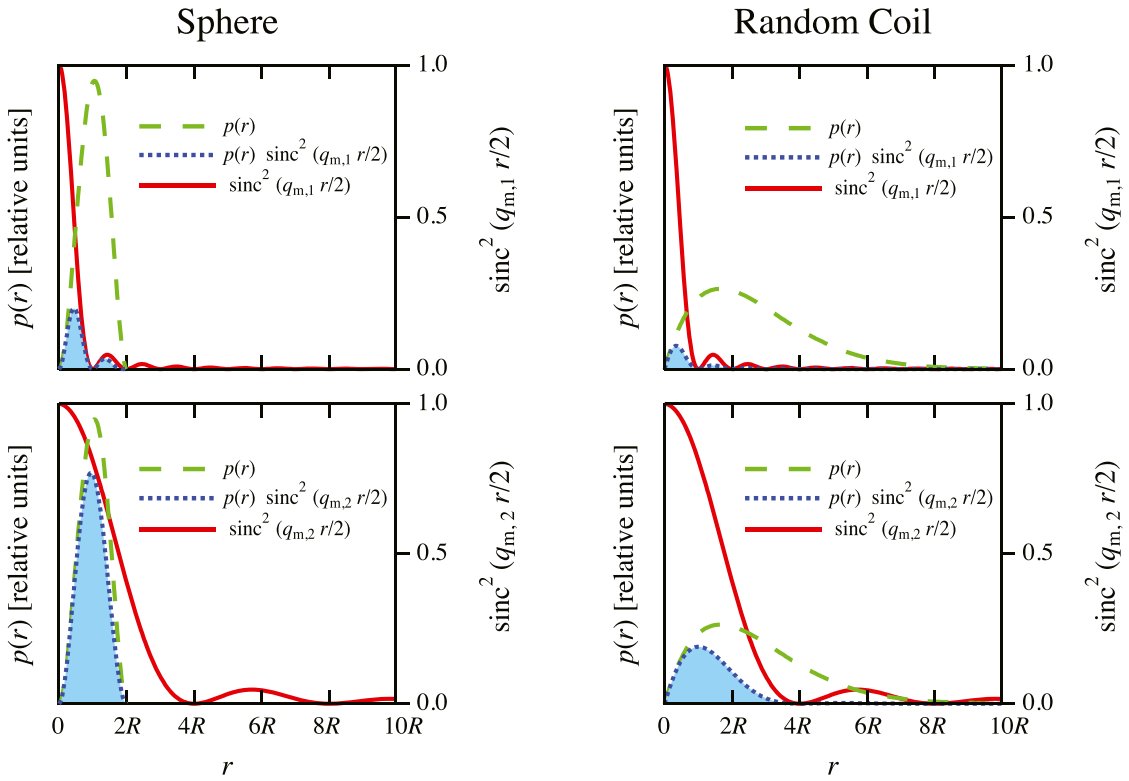

FIGURE 2 A graphical representation of Eq. 4 for a sphere and a random coil at two specific values of q_m. The pair distribution function $p(r)$ for the sphere and a random coil are shown (*left* and *right columns*, respectively). The radius of the sphere is denoted by R. Both objects have the same $I(0)$, which is equal to $\int_0^D p(q)\mathrm{d}r$. (*Top row*) Probe size is equal to R, and $q_{m,1} = 2\pi/R$. (*Bottom row*) Probe size is $4R$, and $q_{m,2} = 2\pi/4R$. The function $\mathrm{sinc}^2(q_m r/2)$ is unitless, and is plotted on a separate axis from $p(r)$. In each panel, $J(q_m)$ is directly proportional to the area (*cyan*) under the dotted curves, and $V_c(q_m)$ is inversely proportional to the area. Due to its lower density and extended shape, the enclosed area for the random coil is smaller than that of the sphere for both probe sizes. As a result, $V_c(q_{m,1})$ for the coil is larger than $V_c(q_{m,1})$ for the sphere. The same holds for $V_c(q_{m,2})$. To see this figure in color, go online.

It has long been theoretically established that integrating $q^2I(q)$ and $qI(q)$ from $q = 0$ to ∞ yields valuable information about a particle's structure (7). However, these integrals frequently do not converge over the $0 < q < q_m^{SAS}$ interval, and are not useful in many practical situations. Within our framework, the upper limit of integration in Eq. 3 can be arbitrary. A finite value of q_m is a strength rather than a weakness, because it provides an adjustable level of resolution for examining a molecule.

Furthermore, our interpretation of $V_c(q_m)$ requires no assumptions regarding compactness, homogeneity, volume, or chord (correlation) length. Assuming a uniform scattering density, Rambo and Tainer (6) used volume and chord length (which they denoted by V_p and l_c, respectively) to write $V_c = V_p/2\pi l_c$. However, both V_p and l_c are functions of the Porod invariant (6,7), which is undefined when $q^2I(q)$ does not converge over the experimentally accessible q-range. This situation frequently occurs for disordered molecules. Even if the Porod invariant could be measured, V_p and l_c represent the ensemble average over all molecular conformations, which is difficult to interpret for noncompact shapes. Our formalism is valid for any molecule and is model-free. In the Results, our theory will be used to interpret measurements of intrinsically disordered proteins, which lack a unique shape.

Effective scaling for globular shapes

For values $q_m^{SAS} = 0.2$–0.5 Å^{-1} corresponding to the typical upper resolution limit of small-angle scattering, the scaling behavior of $V_c(q_m^{SAS})$ lies between the $V_c(q_m) \propto N^0$ and $V_c(q_m) \propto N$ limits discussed above. In this intermediate regime, we use a simple calculation to predict $V_c(q_m^{SAS})$ for globular molecules. In the range $qR_g \lesssim 1$, the Guinier approximation holds for any molecule (7):

$$\frac{I(q)}{I(0)} = \mathrm{Exp}\left(-\frac{q^2R_g^2}{3}\right).$$

This expression may be substituted into Eq. 3. For globular molecules, we assume the majority of the area under the $q\ \mathrm{Exp}[-q^2{R_g}^2/3]$ curve is contained in the interval $0 < q < q_m^{SAS}$, so that the upper limit of the integral may be replaced by infinity. This gives

$$V_c\left(q_m^{SAS}\right) = \frac{2}{3}R_g^2 = \frac{2}{3}R_0^2N^{2/3}, \qquad (6)$$

where the second equality follows from Eq. 1 and a value of $\nu = 1/3$ for a collapsed polymer. Because N is proportional to the total molecular mass M, Eq. 6 also implies $V_c\,(q_m^{SAS}) \propto M^{2/3}$. The predicted value of the coefficient $2R_0{}^2/3$ is less accurate because the Guinier approximation is not valid over the entire range $0 < q < q_m^{SAS}$. Note that the behavior of $V_c(q_m^{SAS})$ qualitatively differs from the $V_c(q_m \rightarrow \infty)$ limit considered in Rambo and Tainer (6) and Svergun et al. (7). As discussed above, this corresponds to the limit of a very small probe size, which instead yields $V_c(q_m \rightarrow \infty) \propto N$. In the Results, we will show that Eq. 6 is in excellent agreement with scattering data for globular proteins.

The above calculation cannot be accurately applied to disordered molecules. For a given value of N, the Guinier approximation is valid over a smaller q-range ($qR_g \lesssim 1$), because disordered molecules exhibit a larger radius of gyration. Consequently, the Guinier approximation becomes qualitatively unreliable over the $0 < q < q_m^{SAS}$ interval. Therefore, the accuracy of Eq. 6 generally decreases with the level of disorder. An analytic result for $V_c(q_m^{SAS})$, in the case of disordered molecules, is beyond the scope of this article.

ANALYSIS DETAILS

The structures of over 9000 globular proteins from the Protein Data Bank (PDB) (12) were analyzed. The proteins were taken from a list compiled by PDB Select (13), whose crystallographic coordinates were determined with an R-factor and resolution less than 0.21. Files containing atoms with identical atomic positions (i.e., two atoms in the same location) were discarded, as well as protein-nucleic acid complexes. Each PDB file was corrected by adding appropriate hydrogen atoms, terminal patches, and disulfide bonds using PSFGEN (14). For each molecule, R_g was calculated using the atomic definition

$$R_g^2 = \frac{\sum_{j,k} b_j b_k \left(\mathbf{r}_j - \mathbf{r}_k\right)^2}{2\left(\sum_j b_j\right)^2}, \qquad (7)$$

where $\mathbf{r}_j$ and b_j are the position and scattering length of atom j, respectively. For each b_j we used the x-ray scattering length at $q = 0$, which is equal to the atomic number. The small-angle x-ray scattering profile of each PDB structure was calculated using the software FoXS (15) with a q-spacing of 10^{-3} Å^{-1}. We found that calculation of R_g and V_c using neutron scattering lengths yielded nearly the same results.

We also compiled measurements of R_g, V_c, and molecular mass from Rambo and Tainer (6). Using small-angle x-ray scattering, they examined 25 globular proteins. Whereas Rambo and Tainer (6) mainly discussed the scaling behavior of $V_c{}^2/R_g$, we analyze the individual data for V_c and R_g.

Using experimental scattering profiles from previous studies, 11 intrinsically disordered proteins (IDPs) were examined as well (16–25). In isolation and under physiological conditions, IDPs lack a stable tertiary structure (26). As opposed to globular proteins, the measured scattering profile of an IDP reflects the average over a large ensemble of conformations. The IDPs were taken from a list compiled by Bernadó and Svergun (27). We only used data that could be reliably extrapolated to $q = 0$, which is required for the profile $I(q)/I(0)$ to be properly normalized. The extrapolation was possible when the data at low q values could be described using the Guinier approximation (i.e., the profile was linear when q^2 was plotted versus $\log[I(q)]$). Table 1 lists the IDPs and their respective values of N, R_g, and V_c. In addition to the IDPs, we also analyzed previously published scattering profiles of Phd_2 (28) and a monoclonal antibody (29).

For all molecules, $V_c(q_m)$ was calculated from Eq. 3 using the trapezoid rule. A script for calculating $V_c(q_m)$ based on experimental data can be downloaded at www.smallangles.net/sassie.

RESULTS

Case studies

Specific examples can be used to illustrate the relationship between molecular structure and $V_c(q_m)$. Figs. 3–5 show the effects of nonspherical shape and disordered chains. In each figure, we also included the unitless Kratky plot (30). In all cases, $V_c(q_m)$ exhibits a hyperbolic shape. When $q_mD \ll 1$, the probe encloses the entire molecule, and $V_c(q_m)$ approaches $2/q_m{}^2$. Outside of the $q_mD \ll 1$ regime, differences in average local structure can be seen. In each case, $V_c(q_m)$ is smaller for the more spherical, compact proteins. While the figures involve molecules with a nearly identical number of residues (N), globular proteins with a specific value of N can be found in the PDB using a customized search.

Fig. 3 compares a Y-shaped antibody (29) with a quasi-spherical globular protein. The antibody consists of three

TABLE 1 Data for the intrinsically disordered proteins shown in Fig. 6. The values of N and R_g were taken directly from references. For all molecules V_c was calculated using a q_m = 0.2 Å^{-1}

Molecule Name	R_g (Å)	V_c (Å^2)	N
MeCP2 (16)	486	62.5	827
Ki-1/57 (17)	292	47.5	660
Pig Calpastatin domain I (18)	148	35.4	290
HrpO (19)	147	35.0	369
II-1 (20)	141	41.0	443
ERM Domain (21)	130	39.6	439
FEZ1 monomer (22)	103	36	365
p53 (1-93) (23)	93	28.7	283
PIR Domain (24)	75	26.5	250
IB5 (20)	73	27.9	229
N-term VS Virus phosphoprotein (25)	68	26	274

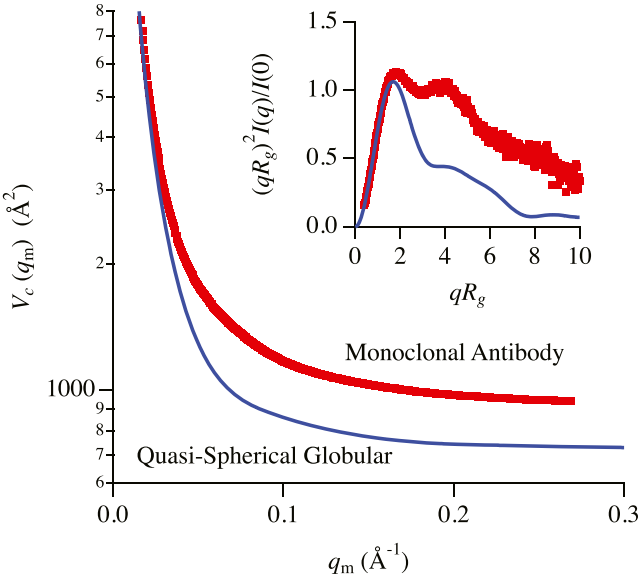

FIGURE 3 Asphericity. $V_c(q_m)$ and the unitless Kratky plot for PDB:4GFI, a quasi-spherical globular protein ($N = 1313$, $R_g = 35.7$ Å), and experimental data for a monoclonal antibody (29) ($N = 1314$, $R_g = 47.5$ Å). To see this figure in color, go online.

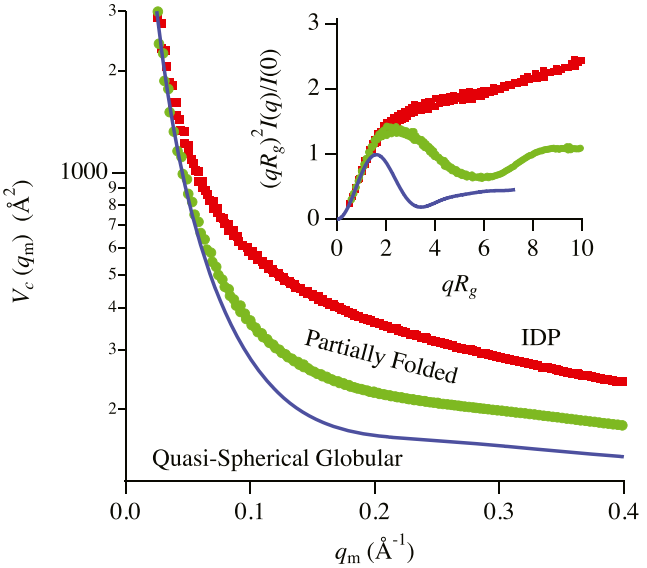

FIGURE 4 Degree of Disorder. $V_c(q_m)$ and the unitless Kratky plot for PDB:2HBG, a compact protein ($N = 148$, $R_g = 14.6$ Å), Phd_2, a partially folded dimer (28) ($N = 146$, $R_g = 22.5$ Å), and HrpO (19), an IDP ($N = 147$, $R_g = 35.0$ Å). The data for Phd_2 and HrpO are based on experimental measurements. To see this figure in color, go online.

compact domains connected by flexible hinges. Due to its asphericity and large conformational ensemble, $V_c(q_m)$ is larger for the antibody.

Fig. 4 contains curves for a globular protein, the partially folded Phd_2 dimer, and an IDP. $V_c(q_m)$ increases with the level of disorder. As shown in the unitless Kratky plot, the curve for the IDP diverges over the experimental q-range, so that the Porod invariant cannot be measured. However, $V_c(q_m)$ remains well defined.

Fig. 5 highlights the effects of disorder and elongation. $V_c(q_m)$ is shown for a quasi-spherical globular protein, an elongated globular protein, and an IDP. Due to its shape, $V_c(q_m)$ for the elongated protein is larger than $V_c(q_m)$ for the spherical protein. In the unitless Kratky plot, this effect corresponds to a shift in the position of the peak. However, compared to $V_c(q_m)$ for the IDP, the differences between the two globular proteins are relatively small. This is due to the average low density of the IDP, compared with the locally compact structure of the globular proteins.

In contrast to Kratky plots, the $V_c(q_m)$ curves allow one to distinguish between compact and disordered molecules at small values of q_m (equivalently, q). In Kratky plots, measurements must extend to a sufficiently large q value to observe the presence or absence of a peak in $q^2I(q)$. The maximum experimental q value is not dictated by the value of R_g alone, because the position of a possible peak in the unitless Kratky plot is not fixed (Figs. 4 and 5). By analyzing $V_c(q_m)$, the disorder of a molecule can be determined at lower values of q_m. In the unitless Kratky plot of Fig. 4, for example, the position of the peak for PDB:2HBG and Phd_2 occurs at $q \approx 0.1$ Å^{-1}. Below that value, the Kratky plot does not provide sufficient information to determine the degree of disorder. In contrast, clear differences in all three molecules can be seen in the $V_c(q_m)$ plot for $q_m < 0.1$ Å^{-1}. Similar effects are present in Figs. 3 and 5 as well. Note also that both $V_c(q_m)$ and Kratky plots distort the molecular features contained within the scattering

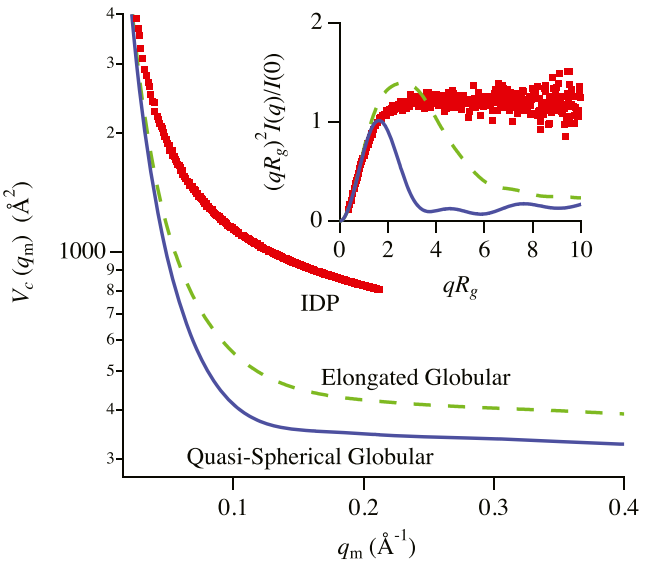

FIGURE 5 Elongation and Disorder. $V_c(q_m)$ and the unitless Kratky plot for three proteins: PDB:3OD3 ($N = 488$, $R_g = 21.6$ Å), a quasi-spherical globular protein; PDB:2JA2 ($N = 487$, $R_g = 31.8$ Å), an elongated globular protein; and MeCP2 (16) ($N = 486$, $R_g = 62.5$ Å), an IDP. The data for MeCP2 are experimental measurements. To see this figure in color, go online.

profile itself, $I(q)$. Combined use of all three plots will allow for a more-comprehensive analysis of scattering data.

Information contained within $V_c(q_m)$ is the most informative when compared with molecules with roughly the same number of residues. Whereas many small-angle scattering studies are devoted to a small number of molecules, globular proteins of equivalent size can usually be found in the PDB using a customized search, which provides a convenient source for comparisons. The theoretical scattering profile of any PDB molecule can be determined using a variety of calculators (see Schneidman-Duhovny et al. (31) for a list), and an integration script for evaluating $V_c(q_m)$ (Eq. 3) can be found at www.smallangles.net/sassie. The profile $V_c(q_m)$ will also be useful for analyzing the disorder of a structure under different solution conditions (6), in which case the number of residues remains unchanged.

Scaling behavior

For general results, we calculated $V_c(q_m)$ and the radius of gyration R_g across many proteins. Unless specified otherwise, we set $q_m = 0.2$ Å^{-1}, a value that can be achieved by nearly all small-angle scattering instruments. For values of $q_m^{SAS} = 0.2$–0.5 Å^{-1} corresponding to the typical upper resolution limit of small-angle scattering, $V_c(q_m^{SAS})$ measures the average local structure on length scales below $\approx$30–10 Å, respectively. $V_c(q_m^{SAS})$ is insensitive to structural properties on length scales that exceed the probe size $2\pi/q_m^{SAS}$. This explains the low level of dispersion when plotting N versus $V_c(q_m^{SAS})$ (Fig. 6). In comparison, R_g (Eq. 7) reflects the average distance between all atom pairs. The dispersion in R_g reflects the diversity in global molecular shapes, whereas the dispersion in $V_c(q_m^{SAS})$ corresponds to density differences on smaller length scales.

Because R_g and V_c have different units, we quantify their dispersion in terms of the average deviation with respect to the number of residues (N) or the molecular mass (M). With R_g, for example, we determine the difference between the actual value of N and the value of N predicted by Eq. 1 using the best-fit parameters for ν and R_0: $N_{pred}^{(k)} = [R_g^{(k)}/R_0]^{1/\nu}$. The average dispersion is defined as

$$\Delta \equiv \frac{1}{\mathcal{N}_{mols}} \sum_{k=1}^{\mathcal{N}_{mols}} \frac{\left| N^{(k)} - N_{pred}^{(k)} \right|}{N_{pred}^{(k)}}, \qquad (8)$$

where $R_g{}^{(k)}$ and $N^{(k)}$ refer to the radius of gyration for molecule k, and number of residues for molecule k, respectively, and $\mathcal{N}_{mols}$ is the number of molecules in the dataset. To measure the dispersion in V_c, we fit the data to Eq. 5 and used $N_{pred}^{(k)} = [V_c{}^{(k)}/\alpha]^{1/\mu}$. Because the data of Rambo and Tainer (6) is listed in terms of molecular mass, we calculated Δ in the same manner as above, but with N replaced by M. Note that in Fig. 6, the symbol "Δ" simply corresponds to the average deviation between each point and the best-fit scaling law along the x axis.

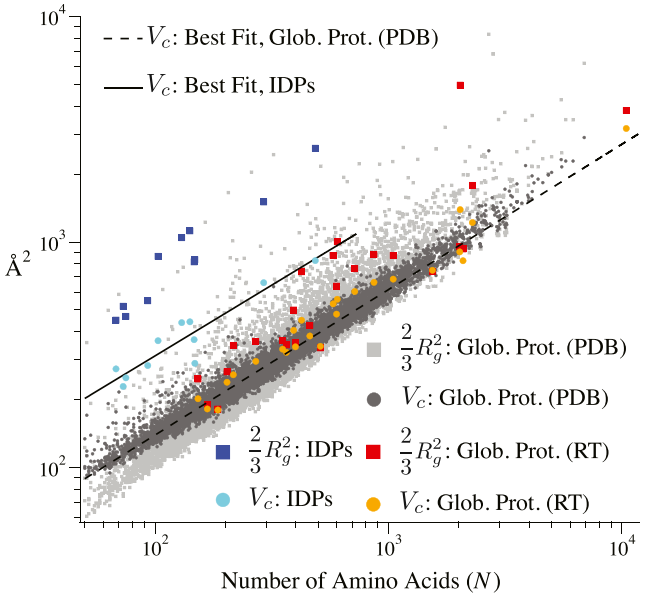

FIGURE 6 $2R_g{}^2/3$ and $V_c[0.2$ Å$^{-1}]$ versus the number of amino acids for 9080 globular proteins from the Protein Data Bank (12) (PDB), 25 globular proteins measured by Rambo and Tainer (6) and 11 intrinsically disordered proteins (IDPs) (16–25). R_g is plotted in this form to coincide with Eq. 6. Best fits to Eq. 5 are also shown. The number of amino acids for the data from Rambo and Tainer (6) was calculated by dividing the total mass of each protein by the average mass per amino acid. See text for details. The data for the PDB proteins are included in the Supporting Material. To see this figure in color, go online.

We fit all three datasets to Eqs. 1 and 5. The results are shown in Table 2. The bootstrap method (32) was used to determine the 95% confidence intervals for our fit parameters. The best-fit values represent the median of the confidence interval, not the mean. As a result, the best-fit parameters do not necessarily lie at the center of the intervals. To obtain fitting parameters for the datasets from Rambo and Tainer (6), we divided the total mass of each molecule by the average mass per residue, $N_{calc}^{(k)} = M^{(k)}/\overline{m}$, where $\overline{m} = 112$ Da. The value of $\overline{m}$ was taken from the PDB dataset. The fitted parameters R_0 and α are affected by the exact value of $\overline{m}$, while the exponents ν and μ are independent of $\overline{m}$.

For both globular protein datasets, the overall relationship between R_g and the number of amino acids is captured by Eq. 1 (see Fig. 6). The best fits for R_0 and ν are shown in Table 2, and coincide with results reported for other globular protein datasets (4,5). They are also consistent the $\nu = 1/3$ prediction for a collapsed polymer (1) (see Hofmann et al. (3) for a discussion on values of R_0 for proteins). The values of R_0 and ν can also be estimated by modeling the proteins as spheres with volume $Nv_{res} = 4\pi R^3/3$, where $v_{res} = 144$ Å^3 is the average approximate volume per residue (33). For a uniform sphere, $R_g{}^2 = 3R^2/5$. From Eq. 1, this yields $R_0 = 2.5$ Å and $\nu = 1/3$, which are close to the

TABLE 2 The scaling behavior of 9080 globular proteins from the Protein Data Bank, 25 globular proteins measured by Rambo and Tainer (6), and 11 intrinsically disordered proteins (16–25)

Molecule type	$\mathcal{N}_{\text{mols}}$	R_0 (Å)	ν[Δ]	α (Å^2)	μ[Δ]
Globular proteins (Protein Data Bank, PDB)	9080	2.4 (2.3–2.4)	0.38[21%] (0.38–0.39)	7.1 (7.1–7.2)	0.65[9.8%] (0.64–0.65)
Globular proteins (Rambo and Tainer (6))	25	3.0[a] (1.5–4.5)	0.36[37%] (0.30–0.47)	7.0[a] (4.9–11)	0.66[16%] (0.60–0.73)
Intrinsically disordered proteins (IDPs)	11	4.4 (2.6–6.2)	0.43[14%] (0.37–0.57)	17 (9.7–34)	0.63[19%] (0.49–0.75)

$\mathcal{N}_{\text{mols}}$ is the number of molecules in each dataset. Data for the number of residues (N) versus R_g and V_c were fit to Eqs. 1 and 5. The best-fit values are listed, along with their 95% confidence intervals, written in parentheses. Due to rounding, some of the best-fit values appear equal to the confidence limits. In brackets, we include the dispersion Δ in the data as defined by Eq. 8.

[a]For the data from Rambo and Tainer (6), the values of R_0 and a were obtained by dividing the molecular mass by the average mass per residue. See text for details.

best-fit values. The $V_c[0.2\ \text{Å}^{-1}] \propto N^{0.65}$ scaling is in excellent agreement with Eq. 6. The dispersion in R_g and V_c is listed in Table 2. The dispersion in V_c is one-half that of R_g, reflecting the compact structure of globular proteins on length scales below $\approx$30 Å.

Globular proteins and IDPs can be clearly distinguished in Fig. 6. At a given value of N, the IDP data for R_g and V_c are always larger than the corresponding R_g and V_c for globular proteins. This reflects the higher level of disorder and lower density of IDPs, and can also be seen in Figs. 4 and 5.

The fit parameters for the IDPs are listed in Table 2. Due to the large dispersion in the data and small sample size, the confidence intervals for the best-fit parameters are large. The fitted values of R_0 and ν are in overall agreement with experimental measurements (3,34). Our calculation based on the Guinier approximation, $V_c(q_{\text{m}}^{\text{SAS}}) = 2{R_g}^2/3$ (Eq. 6), implies that $\mu = 2\nu$. However, this prediction does not apply to disordered molecules, and is not consistent with the best-fit values of μ and ν. Interestingly, the best fit for μ matches the scaling exponent for the globular proteins. However, the large confidence interval makes this result difficult to interpret. The fitted value of α for the IDPs is significantly larger than α for globular proteins, with a small overlap in confidence intervals with the data of Rambo and Tainer (6). The larger value of α corresponds to the offset in V_c between the IDPs and globular proteins in Fig. 6. This offset coincides with the larger values of $V_c[0.2\ \text{Å}^{-1}]$ in Figs. 4 and 5, which reflect the higher level of disorder. Unlike the globular protein datasets, the dispersion in V_c is slightly larger than the dispersion in R_g. The term "intrinsically disordered protein" includes a broad family of molecules, many of which contain both ordered and disordered regions (35). Rather than a scaling law with single values of α and μ, a spectrum of values may be appropriate, with each α and μ corresponding to a different level of disorder (for a specific q_{m}).

Nevertheless, the large offset in V_c between globular proteins and IDPs provides a means to gauge the disorder of a molecule whose V_c has been experimentally measured. A new molecule's (N,V_c) coordinates can be compared with the best-fit scaling curves for IDPs and globular proteins using the parameters in Table 2 (or alternatively, the data itself in Fig. 6). For example, if the (N,V_c) coordinates lie close to the best-fit curve associated with IDPs, it is most likely disordered. If the (N,V_c) location is roughly equidistant between the best-fit curves, it probably contains both compact and disordered regions. However, we stress that the values of α and μ are only based on the measurements of 11 IDPs, and do not constitute a representative sample of IDPs. Scattering data for additional IDPs will certainly be valuable in this regard. Although this method provides a novel approach for measuring the degree of disorder, it will be the most useful when applied in conjunction with other methods as well. Combined analysis of R_g, V_c, $I(q)$, the unitless Kratky plot, and the Porod-Debye plot will allow for more sophisticated investigations of protein structure and flexibility (30,36,37).

Whereas we have focused on the properties of R_g and V_c separately, Rambo and Tainer (6) analyzed products of R_g and V_c. For various integers j and k, they plotted molecular mass versus ${V_c}^j{R_g}^k$, and found that $Q_R \equiv {V_c}^2/R_g$ exhibited the least amount of asymmetry between the data points above and below the fitted lines. Although they gave no explanation for the asymmetry, it mainly originates from the asymmetry in N versus R_g (see Fig. 6). For globular proteins, they found $Q_R \propto M$, which is consistent with our theoretical predictions and the values in Table 2.

Rambo and Tainer (6) also discussed how measurement of Q_R could be used to infer the molecular mass of globular proteins and nucleic acids. Fitting Q_R to a scaling law analogous to Eq. 5, we found that the average error (Δ in Eq. 8) was slightly smaller than that of V_c for the globular protein datasets (see Table 2). The value of Δ not only quantifies the dispersion in the data, but measures the statistical accuracy of the scaling laws. In terms of determining mass, spectrometry and light-scattering techniques are more accurate than the use of average scaling laws. Nevertheless, scaling laws should be convenient for quick-and-dirty measurements as well as high-throughput x-ray analysis (38). Whereas

inferring molecular mass based on V_c or Q_R is relatively accurate for globular proteins, it may not be reliable for all IDPs, which may contain both ordered and disordered regions.

As discussed in the Theory section, the apparent scaling exponent in Eq. 6 approaches 1 as q_m increases. Fitting the PDB globular protein data for $V_c[0.5\ \text{Å}^{-1}]$ gives best-fit values of $\alpha = 4.3\ \text{Å}^2$ and $\mu = 0.71$. Although the exponent is in approximate agreement with Eq. 6, the change in q_m results in fit parameters that significantly differ from those of $V_c[0.2\ \text{Å}^{-1}]$ (Table 2). The value of q_m should therefore be stated explicitly when discussing measurements of V_c and estimating mass.

CONCLUSION

We have demonstrated how the definite integral of the small-angle scattering profile $V_c(q_m)$ can be used to describe the average local structure of any molecule. For a given number of residues, the disorder and shape of any two molecules can be compared by measuring their respective $V_c(q_m)$. The integrated profile of a new molecule can therefore be compared with previous measurements to infer its degree of disorder and/or asphericity. Compared with Kratky plots, analysis of $V_c(q_m)$ provides useful information at smaller q values. While the definite integral effectively washes out specific molecular features contained in $I(q)$, $V_c(q_m)$ provides a measure of the average structure at an adjustable level of resolution. For globular molecules, we have explained the origin of the observed scaling relationship between the integrated scattering profile and the number of residues. The scaling relationship can also be used to estimate the mass of globular proteins. However, the technique is unreliable for determining the mass of IDPs, which cannot be described by a single scaling law.

This work offers a number of future extensions. The scaling relationship for globular proteins should be generalized to include disordered proteins, perhaps by incorporating a polymer form factor (39) (see Eq. 2). Although we mainly compared IDPs with globular proteins, the integrated profile may also be useful for distinguishing between IDPs, since they can sometimes contain compact domains (26). Our general framework for interpreting the integrated profile is not restricted to proteins, and may be applied to other macromolecules as well. Due to its straightforward measurement, the integrated profile should become a standard quantity calculated in all scattering measurements. The integrated profile can be used in concert with other measurements to gain an even deeper understanding of molecular form and function.

SUPPORTING MATERIAL

One spreadsheet is available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(14)00399-3.

We are grateful to Susan Krueger and Nicholas Clark for insightful discussions. We also thank the referees for helpful suggestions.

M.C.W. acknowledges the support of the National Research Council. This work benefitted from CCP-SAS software developed through a joint Engineering and Physical Sciences Research Council (No. EP/K039121/1) and National Science Foundation (No. CHE-1265821) grant.

REFERENCES

1. Doi, M. 1996. Introduction to Polymer Physics. Oxford University Press, London, UK.
2. Kohn, J. E., I. S. Millett, …, K. W. Plaxco. 2004. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. USA.* 101:12491–12496.
3. Hofmann, H., A. Soranno, …, B. Schuler. 2012. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. USA.* 109:16155–16160.
4. Millett, I. S., S. Doniach, and K. W. Plaxco. 2002. Toward a taxonomy of the denatured state: small angle scattering studies of unfolded proteins. *Adv. Protein Chem.* 62:241–262.
5. Dewey, T. 1993. Protein structure and polymer collapse. *J. Phys. Chem.* 98:2250.
6. Rambo, R. P., and J. A. Tainer. 2013. Accurate assessment of mass, models and resolution by small-angle scattering. *Nature.* 496:477–481.
7. Svergun, D., G. Taylor, and L. Feigin. 1987. Structure Analysis by Small-Angle X-Ray and Neutron Scattering. Plenum Press, New York.
8. Mertens, H. D., and D. I. Svergun. 2010. Structural characterization of proteins and complexes using small-angle x-ray solution scattering. *J. Struct. Biol.* 172:128–141.
9. Semenyuk, A. V., and D. I. Svergun. 1991. GNOM—a program package for small-angle scattering data processing. *J. Appl. Cryst.* 24:537–540.
10. Glatter, O. 1977. A new method for the evaluation of small-angle scattering data. *J. Appl. Cryst.* 10:415–421.
11. Svergun, D., and M. Koch. 2003. Small-angle scattering studies of biological macromolecules in solution. *Rep. Prog. Phys.* 66:1735.
12. Berman, H. M., J. Westbrook, …, P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.
13. Joosten, R. P., T. A. te Beek, …, G. Vriend. 2011. A series of PDB-related databases for everyday needs. *Nucleic Acids Res.* 39 (Database issue):D411–D419.
14. Phillips, J. C., R. Braun, …, K. Schulten. 2005. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26:1781–1802.
15. Schneidman-Duhovny, D., M. Hammel, and A. Sali. 2010. FOXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.* 38 (Web Server issue):W540–W544.
16. Yang, C., M. J. van der Woerd, …, K. Luger. 2011. Biophysical analysis and small-angle x-ray scattering-derived structures of MeCP2-nucleosome complexes. *Nucleic Acids Res.* 39:4122–4135.
17. Bressan, G. C., J. C. Silva, …, J. Kobarg. 2008. Human regulatory protein Ki-1/57 has characteristics of an intrinsically unstructured protein. *J. Proteome Res.* 7:4465–4474.
18. Konno, T., N. Tanaka, …, M. Maki. 1997. A circular dichroism study of preferential hydration and alcohol effects on a denatured protein, pig calpastatin domain I. *Biochim. Biophys. Acta.* 1342:73–82.
19. Gazi, A. D., M. Bastaki, …, M. Kokkinidis. 2008. Evidence for a coiled-coil interaction mode of disordered proteins from bacterial type III secretion systems. *J. Biol. Chem.* 283:34062–34068.
20. Boze, H., T. Marlin, …, B. Cabane. 2010. Proline-rich salivary proteins have extended conformations. *Biophys. J.* 99:656–665.
21. Lens, Z., F. Dewitte, …, A. Verger. 2010. Solution structure of the N-terminal transactivation domain of ERM modified by SUMO-1. *Biochem. Biophys. Res. Commun.* 399:104–110.

22. Alborghetti, M. R., A. S. Furlan, …, J. Kobarg. 2010. Human FEZ1 protein forms a disulfide bond mediated dimer: implications for cargo transport. *J. Proteome Res.* 9:4595–4603.

23. Wells, M., H. Tidow, …, A. R. Fersht. 2008. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc. Natl. Acad. Sci. USA.* 105:5762–5767.

24. Moncoq, K., I. Broutin, …, D. Durand. 2004. SAXS study of the PIR domain from the Grb14 molecular adaptor: a natively unfolded protein with a transient structure primer? *Biophys. J.* 87:4056–4064.

25. Leyrat, C., M. R. Jensen, …, M. Jamin. 2011. The N_0-binding region of the vesicular stomatitis virus phosphoprotein is globally disordered but contains transient α-helices. *Protein Sci.* 20:542–556.

26. Eliezer, D. 2009. Biophysical characterization of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* 19:23–30.

27. Bernadó, P., and D. I. Svergun. 2012. Structural analysis of intrinsically disordered proteins by small-angle x-ray scattering. *Mol. Biosyst.* 8:151–167.

28. Garcia-Pino, A., S. Balasubramanian, …, R. Loris. 2010. Allostery and intrinsic disorder mediate transcription regulation by conditional cooperativity. *Cell.* 142:101–111.

29. Clark, N. J., H. Zhang, …, J. E. Curtis. 2013. Small-angle neutron scattering study of a monoclonal antibody using free-energy constraints. *J. Phys. Chem. B.* 117:14029–14038.

30. Durand, D., C. Vivès, …, F. Fieschi. 2010. NADPH oxidase activator $p67^{phox}$ behaves in solution as a multidomain protein with semi-flexible linkers. *J. Struct. Biol.* 169:45–53.

31. Schneidman-Duhovny, D., S. J. Kim, and A. Sali. 2012. Integrative structural modeling with small angle x-ray scattering profiles. *BMC Struct. Biol.* 12:17.

32. Press, W., B. Flannery, …, W. Vetterling. 1990. Numerical Recipes. Cambridge University Press, London, UK.

33. Tsai, J., R. Taylor, …, M. Gerstein. 1999. The packing density in proteins: standard radii and volumes. *J. Mol. Biol.* 290:253–266.

34. Bernadó, P., and M. Blackledge. 2009. A self-consistent description of the conformational behavior of chemically denatured proteins from NMR and small angle scattering. *Biophys. J.* 97:2839–2845.

35. Receveur-Brechot, V., and D. Durand. 2012. How random are intrinsically disordered proteins? A small angle scattering perspective. *Curr. Protein Pept. Sci.* 13:55–75.

36. Rambo, R. P., and J. A. Tainer. 2011. Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers.* 95:559–571.

37. Bernadó, P., E. Mylonas, …, D. I. Svergun. 2007. Structural characterization of flexible proteins using small-angle x-ray scattering. *J. Am. Chem. Soc.* 129:5656–5664.

38. Hura, G. L., A. L. Menon, …, J. A. Tainer. 2009. Robust, high-throughput solution structural analyses by small angle x-ray scattering (SAXS). *Nat. Methods.* 6:606–612.

39. Hammouda, B. 1993. SANS from homogeneous polymer mixtures: a unified overview. *In* Polymer Characteristics Springer, New York, pp. 87–133.