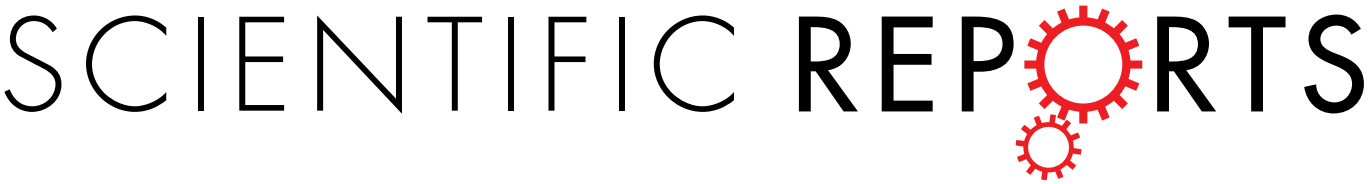

OPEN

A Structure-free Method for Quantifying Conformational Flexibility in proteins

Received: 18 February 2016
Accepted: 08 June 2016
Published: 30 June 2016

Virginia M. Burger[1,*], Daniel J. Arenas[2,*] & Collin M. Stultz[1,3]

All proteins sample a range of conformations at physiologic temperatures and this inherent flexibility enables them to carry out their prescribed functions. A comprehensive understanding of protein function therefore entails a characterization of protein flexibility. Here we describe a novel approach for quantifying a protein's flexibility in solution using small-angle X-ray scattering (SAXS) data. The method calculates an effective entropy that quantifies the diversity of radii of gyration that a protein can adopt in solution and does not require the explicit generation of structural ensembles to garner insights into protein flexibility. Application of this structure-free approach to over 200 experimental datasets demonstrates that the methodology can quantify a protein's disorder as well as the effects of ligand binding on protein flexibility. Such quantitative descriptions of protein flexibility form the basis of a rigorous taxonomy for the description and classification of protein structure.

Thermally induced conformational fluctuations enable proteins to sample a range of structures under physiologic conditions. In many cases, this flexibility is required for a protein to carry out its prescribed function. Quantitative assessments of protein flexibility would therefore further our understanding of the relationship between protein function and structure.

The combination of experiment and computation forms a powerful platform for characterizing protein flexibility. Small-angle X-ray scattering (SAXS), in particular, is one popular experimental method that is often used in this context. Although SAXS typically yields low-resolution information, the combination of SAXS and atomistic simulations can provide insight into conformational changes in proteins and protein flexibility[1–4]. The ensemble optimization method (EOM) and the BILBOMD algorithm, for example, facilitate the construction of conformational ensembles for which the ensemble-averaged theoretical SAXS profile is in agreement with experimentally determined SAXS profiles[3,5]. The resulting conformational ensemble provides a rich dataset that can be used to study the role of protein flexibility in protein function.

Many existing approaches for gaining insight into structural flexibility from experimental data belong to a class of approaches that generate a set of structures to agree with a pre-specified set of experimental observations. This process of generating a set of protein structures that fit a given set of experimental measurements, however, is an underdetermined problem because the number of degrees of freedom in the protein is generally much larger than the number of experimental constraints. While this statement is applicable to all proteins, the problem is most egregious for disordered proteins that, by definition, sample a vast region of conformational space. For these systems there are often many different ensembles that agree with a given set of experimental observations[6,7]. These considerations raise the concern that conclusions arising from these methods may differ depending on the specific choice of the underlying structural model[6]. For example, structural ensembles generated with molecular dynamics simulations can differ depending on the choice of force field and/or solvent model, regardless of whether the protein of interest is disordered or not[8,9]. In addition, while modeling portions of the protein as rigid bodies serves as a useful method for reducing computational time (and is particularly useful for modeling multi-domain proteins)[10–12], it is not always clear what regions of the molecule should, *a priori*, be constrained. The resulting ensemble will therefore depend on the manner in which one chooses to introduce constraints.

[1]Research Laboratory for Electronics, Massachusetts Institute of Technology, 77 Massachusetts Ave. Cambridge MA 02139, USA. [2]Department of Physics, University of North Florida, 1 University of North Fl Dr, Jacksonville, FL 32224, USA. [3]Electrical Engineering and Computer Science & Institute for Medical Engineering and Science Massachusetts Institute of Technology, 77 Massachusetts Ave. Cambridge MA 02139, USA. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to C.M.S. (email: cmstultz@mit.edu)

Consequently, there is a role for structure-free methods that provide information about protein flexibility. Quantitative metrics of protein flexibility calculated from the experimental data alone would facilitate objective comparisons between different proteins, while avoiding the introduction of biases due to the specific choice of structures or simulation protocol. Moreover, metrics that quantify protein flexibility provide a basis for a comprehensive classification scheme for protein structure[13]. Indeed, although proteins are typically categorized as being folded or unfolded, this distinction is overly simplistic because all proteins sample a range of structures at physiologic temperatures. Folded proteins have relatively homogenous ensembles, whereas unfolded proteins have relatively heterogeneous ensembles. Hence, quantitative metrics that provide insight into the heterogeneity within an underlying ensemble would provide a more complete view of the complexity that underlies protein structures and their thermal motions[13].

In this work we describe a new formalism for quantifying protein flexibility from SAXS data. Our approach distinguishes between proteins that have different degrees of disorder and provides novel insights into ligand-induced effects on protein flexibility.

Theory

The Radius of Gyration Distribution (RgD) Model. The measured scattering intensity of a protein is the sum of the scattering intensities of all macromolecular conformations within the protein solution. Thus,

$$I(q) = \int_{\bar{x}} I(q, \bar{x}) P(\bar{x}) d\bar{x}, \tag{1}$$

where q is the magnitude of the scattering vector, $I(q, \bar{x})$ is the scattering intensity of the conformation denoted by $\bar{x} = (x_1, \ldots, x_{3N})$, N is the number of atoms in the protein, and $P(\bar{x})$ is the probability that the macromolecule has conformation $\bar{x}$. Protein flexibility/disorder can be quantified by calculating the entropy, which is a function of $P(\bar{x})$. To compute the probability, $P(\bar{x})$, of any given conformation, the associated Boltzmann factor is required. Unfortunately, determining Boltzmann factors requires knowledge of the exact potential function and modern day empirical potential energy functions are not sufficient for estimating the true density of states under the precise experimental conditions of interest.

To simplify the calculation of the entropy, we propose a model that differentiates conformations based on their radius of gyration, instead of their conformation – a process that reduces the dimensionality of the problem from 3N degrees of freedom to one. Thus, we consider the probabilities of every possible radius of gyration, as opposed to every possible conformation, for our estimation of entropy. The radius of gyration criterion is convenient in SAXS experiments because in the low q region, where the intensity falls off by about one order of magnitude, the intensity is mainly dependent on the size of the macromolecule, i.e. its radius of gyration. We therefore propose a minimalist model in which the intensity profile of a conformation with radius of gyration R_g is represented by the intensity predicted for a sphere with homogeneous charge density[14,15] – a quantity we denote by $I_S(q, R_g)$ and derive in the Supplementary Information.

The scattering intensity in the context of the Radius-of-gyration Distribution (RgD) model, $I_{\mu,\sigma}(q)$, is given by

$$I_{\mu,\sigma}(q) = \int_0^\infty I_S(q, R_g) P_{\mu,\sigma}(R_g) dR_g, \tag{2}$$

where $P_{\mu,\sigma}(R_g)$ is the probability distribution function (pdf) over the different radii of gyration that a protein can adopt in solution. The model uses a log-normal distribution for the pdf,

$$P_{\mu,\sigma}(R_g) = \frac{1}{\sigma R_G \sqrt{2\pi}} e^{\frac{-(\ln(R_g)-\mu)^2}{2\sigma^2}} \tag{3}$$

where μ and σ are the mean and standard deviation of the log-normal distribution. The log-normal distribution has the advantages that it is only defined for positive values of R_g, and $P(R_g)$ approaches zero as R_g approaches zero. For practical use, we set $P(0) = 0$.

To fit the modeled scattering intensity, $I_{\mu,\sigma}(q)$, to the experimental scattering intensity, $I_{\exp}(q)$, we find values of μ and σ that minimize the difference between $I_{\mu,\sigma}(q)$ and $I_{exp}(q)$. More information on the minimization method is provided in the Supplementary Information. The optimal values of μ and σ are denoted as $\hat{\mu}$ and $\hat{\sigma}$. Using these values, the entropy S is computed as:

$$\begin{aligned} S &= -\int_0^\infty P_{\hat{\mu},\hat{\sigma}}(R_g) \ln(P_{\hat{\mu},\hat{\sigma}}(R_g)) dR_g \\ &= \frac{1}{2}(1 + \ln(2\pi\hat{\sigma}^2)) + \hat{\mu}. \end{aligned} \tag{4}$$

A consequence of equation [4] is that the entropy has a lower bound of $-\infty$, a fact that distinguishes it from other discrete entropy measures (e.g., the Shannon entropy) for which the lower bound is zero. The difference in lower bounds between continuous and discrete probability distributions is emphasized by using the term "differential entropy" for the continuous case[16]. The differential entropy expressed by S is a quantitative estimate of the diversity of sampled radii of gyration in solution.

Results and Discussion

RgD on Model Systems. The RgD formalism uses a spherical model to calculate the scattering intensity of a given protein conformation. Modeling protein structures, and conformations within a disordered ensemble, by

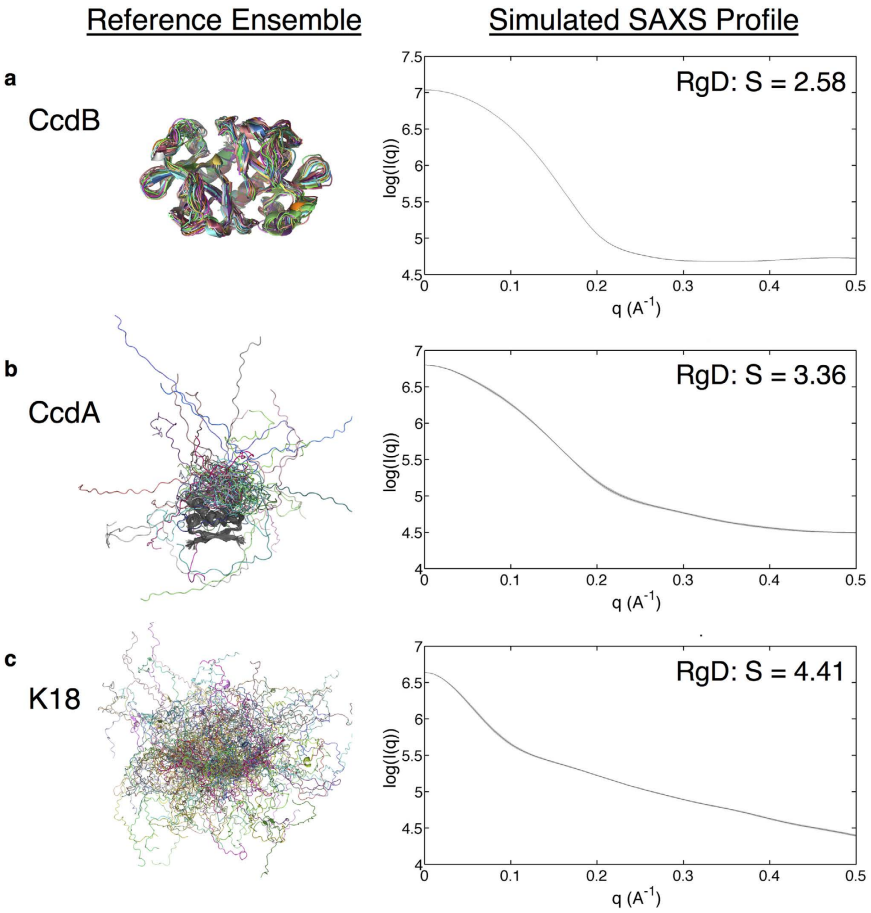

Figure 1. Results from calculations on simulated systems. Alignments of structures in each conformational ensemble are shown on the left. Simulated SAXS profiles and calculated RgD are also shown.

spheres is admittedly a simplification that does not capture the complexity inherent in the structures of biological molecules. However, we were encouraged by the fact that the use of simplified models of biological polymers has a long and rich history of providing important insights into many biological processes[17–25]. To determine whether the RgD formalism has the sensitivity needed to quantify protein flexibility using SAXS intensity profiles, we applied the method to model protein systems representing different degrees of disorder.

We began by choosing three proteins to study – one representing a folded, compact, protein, another a partially disordered protein, and the third an intrinsically disordered protein. Our overall approach was to construct ensembles for each protein, generate a theoretical ensemble average SAXS profile for each ensemble, and then input these data into our RgD algorithm to determine whether the RgD model can produce entropies that are consistent with our understanding of the relative disorder of these systems. In this sense, the constructed conformational ensembles are "reference ensembles", from which experimental observables are calculated. For these simulated experiments the goal is not to generate ensembles that agree with some predefined set of experimental data. By contrast, the structural ensembles represent the "ground truth", which is then used to calculate SAXS profiles. The resulting SAXS profiles are then input to the RgD algorithm to determine whether the method can differentiate proteins according to their flexibility.

For the folded protein we ran molecular dynamics simulations of the 202 residue bacterial toxin protein CcdB from the control of cell death and quiescence gene in *E. coli*[26]. For the partially unfolded protein, we chose the related bacterial antitoxin CcdA, a 144-residue dimer containing a folded core and two intrinsically disordered C-terminal tails, each 34 residues in length[27]. Lastly, for the disordered protein, we used a previously described ensemble for the 130-residue K18 fragment taken from the intrinsically disordered protein tau[6]. Ensemble average SAXS spectra for a given protein were calculated by first computing the individual SAXS spectrum for each structure using Crysol[28] and then averaging the results. These proteins were chosen because they were the focus of prior studies in our group; i.e., structural libraries for these systems already existed. Details of the ensemble construction and calculation of the SAXS profile are discussed in the Supplementary Information.

Results are shown in Fig. 1. The CcdB ensemble contains the least structural heterogeneity and has the lowest RgD entropy (S = 2.58, Fig. 1a). By contrast, the ensemble corresponding to the intrinsically disordered protein, K18, has the highest RgD entropy (S = 4.41, Fig. 1c), and the partially unfolded protein ensemble has an intermediate value (S = 3.36, Fig. 1b). To determine whether these RgD entropy values are significantly different, we estimated the error associated with RgD calculations using the reported errors in experimental scattering intensities

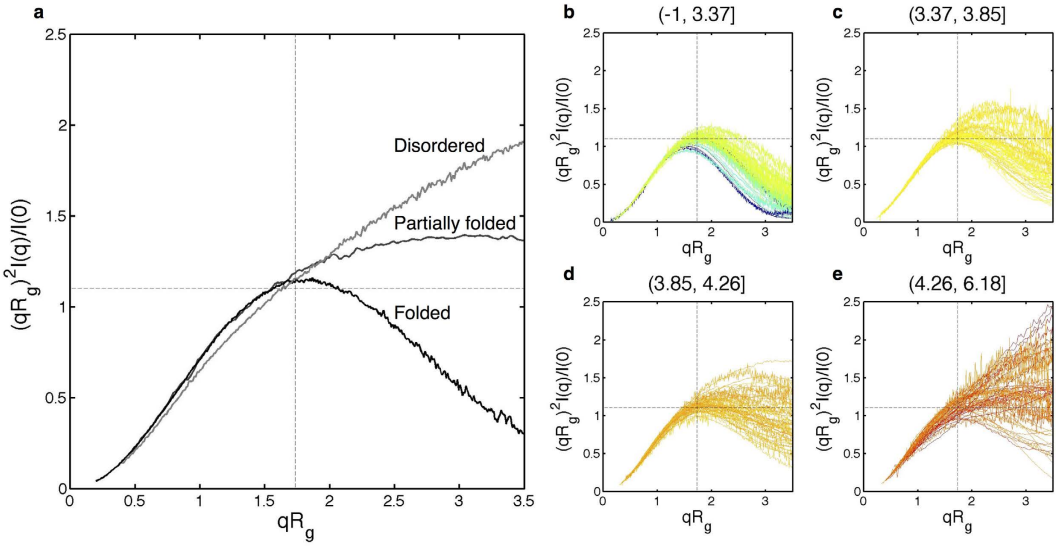

Figure 2. Dimensionless Kratky plots. Dotted lines are drawn at $qR_g = \sqrt{3}$ and $(qR_g)^2I(q)/I(0) = 1.104$. Folded proteins have a local maximum where the two lines intersect. (**a**) Disordered spectrum: C-terminal region of the Bromodomain adjacent to zinc finger protein domain 2B[62]; Partially folded spectrum: Splicing factor U2 Auxiliary Factor 65 KD (U2AF65), residues 148–475[52]; Folded spectrum: Chymotrypsinogen A[63]. (**b–e**) Dimensionless Kratky plots of 226 proteins from the BIOISIS[33] and SASBDB[34] databases organized into quartiles based on their entropy values. The entropy values are divided into four quartiles for the purpose of illustration. The plots are colored such that lower entropies are blue and higher entropies are red.

(see Supplementary Information). In general, RgD entropy errors are less than 1% of the calculated entropy value. These data suggest that differences in the calculated RgD entropies shown in Fig. 1 cannot be attributed to experimental noise alone.

It is important to note that the RgD calculations do not utilize the structural ensembles themselves; i.e., RgD entropies are calculated from the SAXS spectra alone. To determine how our data compare to other structure based estimates that use both the structure and the SAXS profile, we used EOM to compute quantitative estimates of protein flexibility from each of the three model systems we considered[12]. As noted above, EOM takes a SAXS profile as input and generates a corresponding library of structures to arrive at a weighted ensemble of conformations that agree with the SAXS intensity profile. Once the ensemble is determined, the corresponding Shannon entropy provides a measure of the protein's flexibility[12]. In prior applications, this quantity is referred to as R_{flex} and is typically represented as a percentage where 100% represents maximum flexibility[12].

We used the EOM algorithm to generate a pool of 10,000 conformations for each system based on their amino acid sequences and then used these sequence-based conformational pools to fit the theoretical SAXS spectra using the genetic algorithm component of EOM[12]. The predicted R_{flex} values of the selected ensembles – 40% for CcdB, 55% for CcdA and 88% for K18 – showed a range of flexibilities in agreement with the values obtained with the RgD model; i.e., 2.58 for CcdB, 3.36 for CcdA and 4.41 for tau.

While structure-based metrics like R_{flex} are clearly useful for evaluating the flexibility of systems for which the conformational ensemble is unknown, they require the generation of a set of representative structures. Since RgD requires only a SAXS profile to produce an estimate of a system's flexibility, it can provide additional information that may help guide the choice of structural library to use with structure based methods like EOM; e.g., proteins with large RgD entropies should have a large structural library that contains a wide range of different structures, while proteins with small RgD entropies may be better modeled as compact or folded.

Kratky Plots and the RgD Entropy. Kratky plots of SAXS intensity data are commonly used for qualitative assessment of protein disorder. For compact proteins, $I(q)$ will decay as q^{-4}, whereas the scattering intensity of a flexible Gaussian chain will decay as q^{-2} or slower[29]. This suggests that the degree of protein disorder can be inferred from a visual inspection of a plot of $q^2I(q)$ versus q; i.e., a Kratky plot. Compact proteins will have $q^2I(q)$ values that approach zero (or baseline) at high q, while unfolded, or disordered, proteins will generally plateau at intermediate angles followed by continuously increasing values of $q^2I(q)$ at wide angles[1,30,31].

An alternate version of a Kratky analysis renders $(qR_g)^2I(q)/I(0)$ versus qR_g. The x- and y-axes of these plots are dimensionless and therefore are independent of the size and molecular weight of the molecule of interest. Hence these normalized or dimensionless Kratky plots are useful for the analysis of SAXS profiles across different systems. An additional advantage of this formalism is that the dimensionless Kratky plot of a well-folded biopolymer will have a local maximum at $qR_g \approx \sqrt{3}$, which is given by $(qR_g)^2I(q)/I(0) = 3e^{-1} = 1.104$. Homogeneous solutions of folded polymers therefore have dimensionless Kratky plots that have an identifiable characteristic shape[32]. Deviations from this ideal behavior suggest that the macromolecule has conformational flexibility. In Fig. 2a,

characteristic dimensionless Kratky plots for spectra from disordered, partially folded and folded proteins are shown.

To assess how results obtained with the RgD model compare to a Kratky analysis, we calculated entropy values for biopolymers in the BIOISIS database[33] and the Small Angle Scattering Biological Database (SASBDB)[34]. Available entries from either database were excluded from our analysis if: 1) the sample used to obtain the SAXS profile was reported to be aggregated or unpurified; 2) the entry corresponds to unpublished data; or 3) the scattering profile only sampled q values less than 0.3 $Å^{-1}$. This latter requirement ensured that each entry had enough data to perform a meaningful analysis using standard approaches such as a Kratky plot. This screen left a total of 226 experimental datasets for our analysis (Supplementary Tables S1 and S2).

Figure 2b–e show normalized Kratky plots for the datasets in our analysis, divided into four quartiles according to the entropy (S) computed by RgD. The entropy values are divided into four quartiles for the purpose of illustration only. Entropy values vary between −1 and 6.18, where entries that fall in the lowest quartile ($S \leq 3.37$) have dimensionless Kratky plots that are characteristic of compact, folded, states (Fig. 2b). By contrast, dimensionless Kratky plots in the highest quartile ($S > 4.26$) are characteristic of flexible or disordered biopolymers (Fig. 2e). Entropy values between 3.37 and 4.26 correspond to intermediate behavior, with values between 3.86 and 4.27 associated with relatively increased flexibility (Fig. 2c,d).

It is important to recognize that the RgD model was not designed to simply quantify the information contained in Kratky plots. Indeed, since Kratky plots can be difficult to interpret and are sometimes unable to provide an accurate assessment of protein flexibility[35,36], a simple reproduction of insights obtained from a Kratky analysis should not, in and of itself, be the sole metric of success[37]. To demonstrate that the model provides information that is distinct, and complementary, to existing SAXS based methods for the assessment of protein flexibility, we used the model to quantify ligand-induced changes in protein flexibility.

The RgD Entropy and Ligand-Binding. We began by searching the BIOISIS database to find a suitable subset of protein-ligand complexes for additional analyses[38]. Only entries where both the spectra of the free and complexed protein were obtained by the same research group, and under similar experimental conditions, were considered. Below we discuss our results below, in light of the available experimental data.

MnmE. *E. coli* MnmE plays a crucial role in modifying wobble uridine in tRNA[39]. In separate studies, X-ray crystallography, electron paramagnetic resonance (EPR), and SAXS experiments were used to study the structure of MnmE in 1) the free state, 2) bound to the transition state analogue GDP-A1Fx, and 3) bound to the ground state analogue GppNHp[40,41]. In the free state MnmE adopts an open structure where two of its domains (the G-domains) are separated, while binding to GDP-A1Fx causes the protein to adopt a "closed" conformation where the G-domains dimerize[41]. By contrast, binding to GppNHp induces the protein to adopt a mixture of closed and open conformations, where approximately 88% of the protein is in the closed state and 12% is in the open state[40].

Dimensionless Kratky representations of the three systems are very similar in that all three proteins have a local maximum at $qR_g \approx \sqrt{3}$, and at this value $(qR_g)^2 I(q)/I(0) = 3e^{-1} = 1.104$ (Fig. 3a). It is therefore difficult to make any conclusions about the relative stability of these complexes from these data alone. Given that the dimensionless Kratky plots provide little, if any, insight into ligand-induced changes in protein flexibility, we performed a Porod-Debye analysis to determine how this approach compares to the RgD model. The Porod-Debye relationship dictates that for a compact polymer the scattering intensity decays as q^{-4} and that for some small range of q, a plot of $q^4 I(q)$ vs. q^4 will achieve a plateau, which is a function of the molecule's surface area and its electron density contrast with respect to the surrounding solvent[37,42,43]. In practice, the range of q where the Porod-Debye law is applicable – the Porod-Debye region – is estimated from the position of the first peak in the corresponding Porod plot (i.e., $q^4 I(q)$ vs q). Proteins that have considerable flexibility decay slower than q^{-4} and therefore do not reach a plateau in the Porod-Debye region.

A Porod-Debye analysis does clarify the role of flexibility to some degree. The unbound protein does not have a clear Porod-Debye plateau (Fig. 3b, black), while the bound proteins do (Fig. 3b, green and purple). These data suggest that binding of both GDP-A1Fx and GppNHp reduces MnmE flexibility. However, it is not clear from these data which analog causes the greatest reduction in flexibility after binding. Without additional information it is difficult to make conclusive statements about relative protein flexibility from these observations.

The RgD model suggests that binding of both the ground state analog and the transition state analog reduces the flexibility of the protein and that binding of the transition state analog, GDP-A1Fx, is associated with the greatest reduction in flexibility (Fig. 3a). Moreover, as errors in the experimental scattering intensities correspond to small errors in the calculated RgD entropy values (approximately 0.03 for free MnmE and 0.01 for MmmE bournd to GppNHp, and 0.003 for MnmE bound to GDP-A1Fx, see Supplementary Table S3), it is difficult to ascribe the differences in RgD values between the three systems to experimental error alone.

Since GDP-A1Fx binding causes the protein to adopt a closed state, these observations argue that the closed state is the most rigid. The fact that the MnmE-GppNHp complex has an intermediate value for the entropy is consistent with the observation that GppNHp binding leads to an equilibrium distribution of closed and open states[40,41].

wtTIA-1 RRM123. T-cell intracellular antigen-1 (wtTIA-1) plays a crucial role in pre-mRNA splicing and is an important regulator of translation[44]. It contains three RNA recognition motifs (RRMs) that bind U-rich RNA segments downstream of other weak splice sites. Recently the binding of all three RRMs (wtTIA-1 RRM123) to U-rich RNA sequences was studied using SAXS and isothermal titration calorimetry (ITC)[45,46].

Dimensionless Kratky plots of wtTIA-1 RRM123 in its free and bound state suggest that binding is associated with a loss of protein flexibility. The Kratky plot for the bound state (Fig. 3c, green) has a local maximum, which

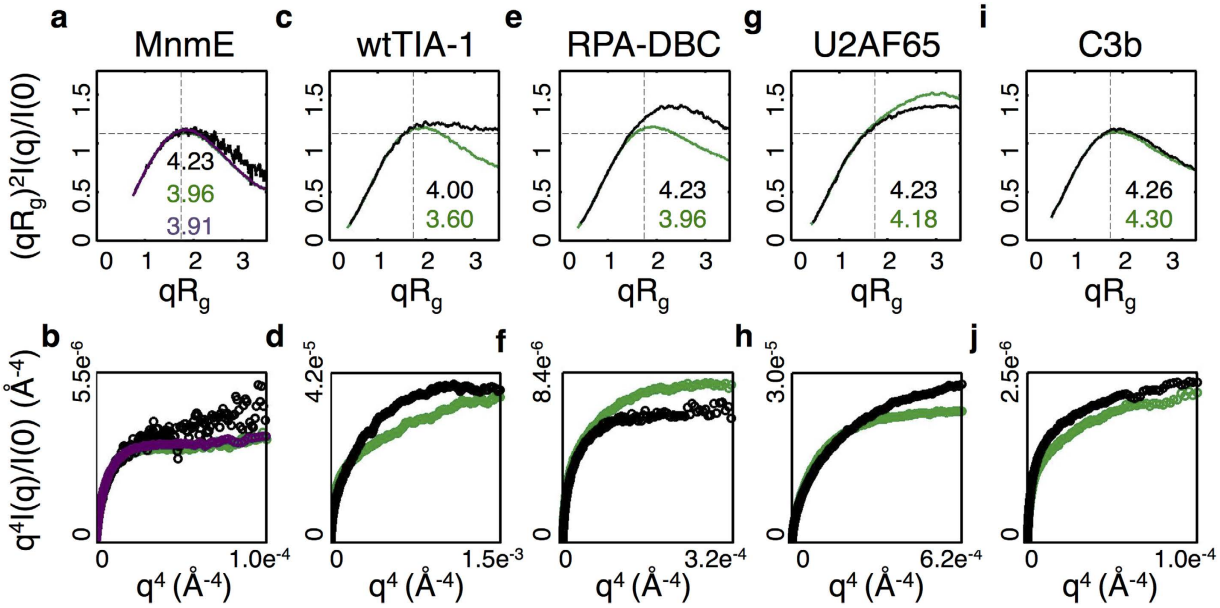

Figure 3. Dimensionless Kratky plots (top row), calculated RgD entropy values (insets in top row), and Porod Debye plots (bottom row) for MnmE: E. coli MnmE in isolation (black) and bound to GppNHp (green), and GDP-AlFx (purple); wtTIA-1: The alterative splicing factor wtTIA-1 RRM123 in the absence of RNA (black) and bound to 11-nucleotide AU-rich segment taken from the 3′-untranslated region of tnf-α (green); RPA-DBC: The DNA-binding core of heterotrimeric Replication protein A in the absence (black) and presence (green) of a 30 nucleotide ssDNA substrate; U2AF65: U2 auxiliary factor residues 148-475, in the absence (black) and presence of RNA (green); C3b: Complement fragment C3b in the unbound (black) state and bound to the extracellular fibrinogen binding protein (Efb) from S. aureus (green). Since we work with normalized Intensity profiles (that are divided by *I(0)*) the y-axis of each Porod-Debye plots is divided by *I(0)*.

is close to the ideal value for a folded polymer, relative to the plot corresponding to the unbound state (Fig. 3c, black). However, Porod-Debye plots of wtTIA-1 RRM123 yield contradictory information (Fig. 2d, black). While the free protein reaches a clear Porod-Debye plateau by $q^4 \approx 0.18^4 = 0.001$ Å^4, the plateau is lost in the bound state (Fig. 2d, green). A plot of $q^3I(q)^3$ vs. q^3 for the bound state further demonstrates that $I(q)$ decays as q^{-3} in the Porod-Debye region instead of the expected q^{-4} for a compact polymer, thereby suggesting that binding makes the protein more flexible (Supplementary Fig. S1)[37].

ITC studies suggest that RNA binding to wtTIA-1 is associated with large unfavorable changes in the binding entropy (approximately 30 kcal/mol)[46]. In general, the total binding entropy is a function of several different physical phenomena including, for example, dynamical changes in the binding species, release of ordered water molecules, and the vibrational spectra of both the bound and unbound states[47]. The RgD model suggests that RNA binding is associated with a decrease in the entropy (Fig. 3c), and therefore argues that a decrease in conformational entropy contributes to the large unfavorable entropic contribution to the binding energy. While there is certainly precedent for ligand binding to increase the conformational entropy of a protein[48,49], as the Porod-Debye plots suggest, the large unfavorable entropy associated with RNA binding is more consistent with a loss of protein flexibility[46,50], as the RgD model suggests.

RPA-DBC. Replication protein A (RPA) is multi-domain protein that plays an important role in regulating DNA processing. Recently a combination of SAXS and molecular dynamics simulations was used to study binding of the DNA-binding core of RPA (RPA-DBC) to a 30-nucleotide ssDNA substrate[51]. Extensive simulations were performed to generate structures that were consistent with experimentally determined SAXS profiles of the free and bound protein. A conformational analysis of the resulting ensembles suggested that RPA-DBC bound to ssDNA is more compact relative to the free protein and that the bound state samples a smaller range of radii of gyration relative to the unbound protein. These observations are echoed by our calculations in that binding to DNA leads to a decrease in the RgD entropy (Fig. 3e). Since the RgD model quantifies the diversity of sampled radii of gyration, a decrease in the RgD entropy means that the bound state samples a smaller range of radii of gyration in solution.

The dimensionless Kratky plots are also consistent with these data in that the plot of the bound protein has a peak located at the ideal position for a folded protein, whereas the free protein does not (Fig. 3e). A Porod-Debye plot of the bound complex has a clear plateau (Fig. 3f, green), and at first glance a similar plot for the free protein plateaus as well, albeit to a lower value (Fig. 3f, black). The fact that both plots plateau to different values suggests that the free and bound structures have different spectroscopic properties. Since flexibility cannot be inferred from the value of the plateau itself, it is unclear how these observations relate to any changes in protein flexibility[37]. It could be argued that the Porod-Debye plot of the free protein slowly increases at relatively wide angles

($q^4 > 0.00025 \approx 0.125^4$ Å^4, Fig. 3f, black), but this may be secondary to experimental noise (or poor buffer subtraction) – phenomena that may be seen at higher q values[31]. Indeed, at high-q the scattering profile of the free protein has larger variations than that of the bound complex (Supplementary Fig. S2). In short, it is difficult to reconcile observations arising from this Porod-Debye analysis with the results of the combined SAXS/simulation study mentioned above. In this regard, the RgD model provides clarifying information that complements the results of the Kratky and Porod-Debye analyses.

U2AF65. The splicing factor U2AF65 assembles on RNA during the early stages of pre-mRNA splicing. During assembly U2AF65 binds to pre-mRNA at the 3′ splice site. Recently the binding of the SF1/U2AF65 Splicing Factor Complex was studied using SAXS[52]. Experiments with U2A65 utilized a construct (residues 148–475) containing three domains: one that recognizes the N-terminal region of splicing factor 1; and two RNA recognition domains, each of which bind RNA[53]. Dimensionless Kratky plots of U2AF65 suggest that both the unbound and bound states are flexible (Fig. 3g). Given that the individual domains are known to be folded, these data are consistent with U2AF65 being composed of folded modular domains that are connected by flexible linkers[52]. Nevertheless, it is difficult to make definitive statements about the relative flexibility of the bound state from these data alone. The RgD model predicts that binding leads to a decrease in the system entropy (Fig. 3g). However, it should be mentioned that the decrease is small and very close to the errors in entropy that we estimated using noise simulations (see Supplementary Table S1). A Porod-Debye plot of the bound state of the U2AF65 spectrum has a plateau (Fig. 3h, green) relative to its free state (Fig. 3h, black), suggesting that binding results in a decrease in the system entropy, a finding consistent with the RgD results.

C3b. The complement fragment C3b plays an important role in human immunity[54]. Interactions of C3b trigger a host of inflammatory responses that eventually lead to the death of foreign microorganisms. Binding of C3b to the extracellular fibrinogen-binding (Efb) protein from *S. aureus* was recently studied using a combination of SAXS and molecular modeling[55]. Dynamical simulations of C3b were conducted to generate a minimal set of conformers that agreed with SAXS profiles of the protein in its free and bound forms. The resulting ensembles suggest that C3b samples both open and closed states in its unbound form. In the open state, two domains of C3b (the CUB and TED domains), which are connected to the core of the protein via a flexible linker, adopt conformations that are separated from the core. By contrast, in the closed state, the CUB-TED domains are packed against the protein core. A combination of hydrogen-deuterium exchange experiments and molecular simulations suggest that Efb binds at the interface between the TED domain and protein core, and that Efb binding stabilizes the protein in the open state[55].

Dimensionless Kratky plots of the free and bound protein are very similar (Fig. 3i) and the associated Porod-Debye plots do not plateau, making it unclear whether binding has any influence on flexibility (Fig. 3j). The RgD entropy calculations suggest that both the free and bound proteins are very flexible in that their RgD entropy values place them in the third and fourth quartiles of proteins in the BIOISIS and SASBDB databases (Figs 3i and 2d,e). Moreover, the calculated entropy for the bound state is larger than the entropy of the free protein, suggesting that the bound protein is more flexible than the unbound protein. However, it should be noted that the difference between these values are quite small and within the range of error associated with RgD calculations (n.b. the errors associated with RgD calculations on C3b are 0.01, as shown in Supplementary Table S3). Since binding of Efb stabilizes the open state, these calculations suggest that the bound, and predominantly open, state is able to sample a range of radii of gyration that is similar to, or possibly larger than, that of the unbound protein.

The aforementioned simulations argue that the free protein samples closed and open states that have similar radii of gyration and that the measured radius of gyration of the free protein is a weighted sum over these values[55]. Similarly, the RgD entropy, which is calculated from the RgD model, is also a weighted sum of entropic contributions from both the closed and open states. If the open state were more flexible than the closed state, then stabilization of the open state through binding by Efb would result in an increase in the overall entropy. The dynamical simulations mentioned above utilized a protocol where the CUB-TED domains were modeled as rigid bodies connected by flexible linkers, with the rest of the protein held in a fixed position. In light of this, it is difficult to gauge the relative flexibilities of the open and closed states, and how binding affects the flexibility of the open state, from these calculations. Nonetheless, the entropy computed for the bound and unbound SAXS profiles with RgD allows us to predict that Efb binding to the open states results in the protein sampling a wider range of radii of gyration.

Conclusions

A number of experimentally derived metrics have been developed to quantify protein flexibility. For example, quantitative metrics that facilitate the study of protein flexibility include X-ray diffraction at different temperatures[56], NMR relaxation experiments[57,58], and atomic force microscopy[59]. These approaches, however, often require experimental conditions that are quite different from the solution state, or the use of isotopically labeled protein. In addition, these experiments only account for motions that occur on the microsecond-to-millisecond time scales. SAXS, albeit a low-resolution technique, has the advantage that it provides information about the structure of the protein in solution without the use of special isotopes, and also provides information about large conformational changes that typically occur on long time scales[3,60].

Our approach estimates the pdf over the different radii of gyration that a biomolecule can adopt in solution using the SAXS profile alone. Once the pdf is known, the entropy can be calculated in a straightforward manner. Since the entropy computed by RgD quantifies the diversity of radii of gyration sampled by a protein in solution, this method provides a direct measure of a system's disorder. Application to over 200 proteins in the BIOISIS[33] and SASBDB[34] databases demonstrates that the RgD model can provide information about the degree of a protein's disorder, as well as insight into how ligand binding affects protein flexibility.

The RgD entropy is a continuous parameter that quantifies the extent of disorder in a protein's conformational ensemble; i.e., the set of thermally accessible conformations available in solution. It is our view that such quantitative descriptions of protein structure are more accurate than the traditional binary terms, "folded" and "unfolded", which are often used to classify proteins. Indeed, a more accurate description of protein structure should entail a characterization of the heterogeneity within a protein's conformational ensemble[13]. The importance of this realization is highlighted by the fact that not all folded proteins are created equal. Some "folded" ensembles are more heterogeneous than others, as evidenced by the range of RgD entropies that are observed for different folded proteins (Fig. 2b). Similarly, disordered proteins often exhibit preferences for particular structural features[61]. These considerations reinforce the notion that quantitative metrics describing the heterogeneity within a protein's ensemble provide a more comprehensive assessment of protein structure than binary classification.

References

1. Mertens, H. D. & Svergun, D. I. Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *J Struct Biol* **172,** 128–141, doi: 10.1016/j.jsb.2010.06.012 (2010).
2. Rambo, R. P. & Tainer, J. A. Bridging the solution divide: comprehensive structural analyses of dynamic RNA, DNA, and protein assemblies by small-angle X-ray scattering. *Current Opinion in Structural Biology* **20,** 128–137, doi: http://dx.doi.org/10.1016/j.sbi.2009.12.015 (2010).
3. Bernado, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc* **129,** 5656–5664, doi: 10.1021/ja069124n (2007).
4. Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.* **40,** 191–285 (2007).
5. Pelikan, M., Hura, G. L. & Hammel, M. Structure and flexibility within proteins as identified through small angle X-ray scattering. *General Physiology and Biophysics* **28,** 174–189, doi: 10.4149/gpb_2009_02_174 (2009).
6. Fisher, C. K., Huang, A. & Stultz, C. M. Modeling intrinsically disordered proteins with Bayesian statistics. *J Am Chem Soc* **132,** 14919–14927, doi: 10.1021/ja105832g (2010).
7. Ganguly, D. & Chen, J. Structural interpretation of paramagnetic relaxation enhancement-derived distances for disordered protein states. *J Mol Biol* **390,** 467–477, doi: 10.1016/j.jmb.2009.05.019 (2009).
8. Henriques, J., Cragnell, C. & Skepö, M. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *Journal of Chemical Theory and Computation* **11,** 3420–3431, doi: 10.1021/ct501178z (2015).
9. Lindorff-Larsen, K. *et al.* Systematic validation of protein force fields against experimental data. *PLoS One* **7,** e32131, doi: 10.1371/journal.pone.0032131 (2012).
10. Petoukhov, M. V. & Svergun, D. I. Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophysical journal* **89,** 1237–1250 (2005).
11. Saw, W. G. *et al.* Structural insight and flexible features of NS5 proteins from all four serotypes of Dengue virus in solution. *Acta crystallographica Section D, Biological crystallography* **71,** 2309–2327 (2015).
12. Tria, G., Mertens, H. D. T., Kachala, M. & Svergun, D. I. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ* **2,** 207–217 (2015).
13. Fisher, C. K. & Stultz, C. M. Protein structure along the order-disorder continuum. *J Am Chem Soc* **133,** 10022–10025, doi: 10.1021/ja203075p (2011).
14. Stein, R. S., Wilson, P. R. & Stidham, S. N. Scattering of Light by Heterogeneous Spheres. *Journal of Applied Physics* **34,** 46–50, doi: http://dx.doi.org/10.1063/1.1729087 (1963).
15. Fournet, G. & Guinier, A. Small-Angle Scattering of X-rays. (John Wiley & Sons, Inc. (NY), Chapman & Hall, Ltd. (London), 1955).
16. Beirlant, J., Dudewicz, E. J., Györfi, L. & Van der Meulen, E. C. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences* **6,** 17–39 (1997).
17. Mirny, L. & Shakhnovich, E. Protein folding theory: from lattice to all-atom models. *Annu Rev Biophys Biomol Struct* **30,** 361–396, doi: 10.1146/annurev.biophys.30.1.361 (2001).
18. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The protein folding problem. *Annu Rev Biophys* **37,** 289–316, doi: 10.1146/annurev.biophys.37.092707.153558 (2008).
19. Tozzini, V. Coarse-grained models for proteins. *Curr Opin Struct Biol* **15,** 144–150, doi: 10.1016/j.sbi.2005.02.005 (2005).
20. Aaron, B. & Gosline, J. Elastin as a random-network elastomer: A mechanical and optical analysis of single elastin fibers. *Biopolymers* **20,** 1247–1260 (1981).
21. Anfinsen, C. B. & Scheraga, H. A. Experimental and theoretical aspects of protein folding. *Adv Protein Chem* **29,** 205–300 (1975).
22. Levitt, M. Molecular dynamics of native protein. I. Computer simulation of trajectories. *Journal of molecular biology* **168,** 595–617 (1983).
23. McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of folded proteins. *Nature* **267,** 585–590 (1977).
24. Shakhnovich, E. I. Theoretical studies of protein-folding thermodynamics and kinetics. *Current opinion in structural biology* **7,** 29–40 (1997).
25. Taketomi, H., Ueda, Y. & Go, N. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *International journal of peptide and protein research* **7,** 445–459 (1975).
26. De Jonge, N. *et al.* Rejuvenation of CcdB-Poisoned Gyrase by an Intrinsically Disordered Protein Domain. *Molecular Cell* **35,** 154–163, doi: 10.1016/j.molcel.2009.05.025 (2009).
27. Madl, T. *et al.* Structural basis for nucleic acid and toxin recognition of the bacterial antitoxin CcdA. *Journal of molecular biology* **364,** 170–185 (2006).
28. Svergun, D., Barberato, C. & Koch, M. H. J. CRYSOL - a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *Journal of Applied Crystallography* **28,** 768–773, doi: 10.1107/S0021889895007047 (1995).
29. Bernado, P. & Svergun, D. I. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Molecular Biosystems* **8,** 151–167, doi: 10.1039/c1mb05275f (2012).
30. Debye, P. Molecular-weight determination by light scattering. *J Phys Colloid Chem* **51,** 18–32 (1947).
31. Hammel, M. Validation of macromolecular flexibility in solution by small-angle X-ray scattering (SAXS). *European Biophysics Journal* **41,** 789–799, doi: 10.1007/s00249-012-0820-x (2012).
32. Receveur-Brechot, V. & Durand, D. How random are intrinsically disordered proteins? A small angle scattering perspective. *Curr Protein Pept Sci* **13,** 55–75 (2012).
33. Hura, G. L. *et al.* Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat Meth* **6,** 606–612, doi: http://www.nature.com/nmeth/journal/v6/n8/suppinfo/nmeth.1353_S1.html (2009).
34. Valentini, E., Kikhney, A. G., Previtali, G., Jeffries, C. M. & Svergun, D. I. SASBDB, a repository for biological small-angle scattering data. *Nucleic acids research* **43,** D357–D363 (2015).

35. Uversky, V. N. *et al.* Natively unfolded human prothymosin alpha adopts partially folded collapsed conformation at acidic pH. *Biochemistry* **38,** 15009–15016 (1999).
36. Williams, G. J. *et al.* ABC ATPase signature helices in Rad50 link nucleotide state to Mre11 interface for DNA repair. *Nature structural & molecular biology* **18,** 423–431 (2011).
37. Rambo, R. P. & Tainer, J. A. Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers* **95,** 559–571, doi: 10.1002/bip.21638 (2011).
38. Valentini, E., Kikhney, A. G., Previtali, G., Jeffries, C. M. & Svergun, D. I. SASBDB, a repository for biological small-angle scattering data. *Nucleic acids research*, gku1047 (2014).
39. Yim, L. *et al.* The GTPase activity and C-terminal cysteine of the Escherichia coli MnmE protein are essential for its tRNA modifying function. *The Journal of biological chemistry* **278,** 28378–28387 (2003).
40. Fislage, M. *et al.* SAXS analysis of the tRNA-modifying enzyme complex MnmE/MnmG reveals a novel interaction mode and GTP-induced oligomerization. *Nucleic Acids Research* **42,** 5978–5992, doi: 10.1093/nar/gku213 (2014).
41. Meyer, S. *et al.* Kissing G Domains of MnmE Monitored by X-Ray Crystallography and Pulse Electron Paramagnetic Resonance Spectroscopy. *PLoS Biol* **7,** e1000212, doi: 10.1371/journal.pbio.1000212 (2009).
42. Porod, G. Die Rontgenkleinwinkelstreuung Von Dichtgepackten Kolloiden Systemen .1. *Kolloid-Zeitschrift and Zeitschrift Fur Polymere* **124,** 83–114, doi: 10.1007/Bf01512792 (1951).
43. Debye, P., Anderson, H. R. & Brumberger, H. Scattering by an Inhomogeneous Solid .2. The Correlation Function and Its Application. *Journal of Applied Physics* **28,** 679–683, doi: Doi 10.1063/1.1722830 (1957).
44. Forch, P. *et al.* The apoptosis-promoting factor TIA-1 is a regulator of alternative pre-mRNA splicing. *Mol Cell* **6,** 1089–1098 (2000).
45. Bauer, W. J., Heath, J., Jenkins, J. L. & Kielkopf, C. L. Three RNA Recognition Motifs Participate in RNA Recognition and Structural Organization by the Pro-Apoptotic Factor TIA-1. *Journal of Molecular Biology* **415,** 727–740, doi: 10.1016/j.jmb.2011.11.040 (2012).
46. McLaughlin, K. J., Jenkins, J. L. & Kielkopf, C. L. Large Favorable Enthalpy Changes Drive Specific RNA Recognition by RNA Recognition Motif Proteins. *Biochemistry* **50,** 1429–1431, doi: 10.1021/bi102057m (2011).
47. Zhou, H. X. & Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chem. Rev.* **109,** 4092–4107 (2009).
48. Zidek, L., Novotny, M. V. & Stone, M. J. Increased protein backbone conformational entropy upon hydrophobic ligand binding. *Nat Struct Mol Biol* **6,** 1118–1121 (1999).
49. Diehl, C. *et al.* Protein Flexibility and Conformational Entropy in Ligand Design Targeting the Carbohydrate Recognition Domain of Galectin-3. *Journal of the American Chemical Society* **132,** 14577–14589, doi: 10.1021/ja105852y (2010).
50. Chao, J. A., Prasad, G. S., White, S. A., Stout, C. D. & Williamson, J. R. Inherent Protein Structural Flexibility at the RNA-binding Interface of L30e. *Journal of Molecular Biology* **326,** 999–1004, doi: http://dx.doi.org/10.1016/S0022-2836(02)01476-6 (2003).
51. Brosey, C. A. *et al.* A new structural framework for integrating replication protein A into DNA processing machinery. *Nucleic Acids Research* **41,** 2313–2327, doi: 10.1093/nar/gks1332 (2013).
52. Gupta, A., Jenkins, J. L. & Kielkopf, C. L. RNA Induces Conformational Changes in the SF1/U2AF(65) Splicing Factor Complex. *Journal of molecular biology* **405,** 1128–1138, doi: 10.1016/j.jmb.2010.11.054 (2011).
53. Zamore, P. D., Patton, J. G. & Green, M. R. Cloning and Domain-Structure of the Mammalian Splicing Factor U2af. *Nature* **355,** 609–614, doi: 10.1038/355609a0 (1992).
54. Ricklin, D., Hajishengallis, G., Yang, K. & Lambris, J. D. Complement: a key system for immune surveillance and homeostasis. *Nature immunology* **11,** 785–797 (2010).
55. Chen, H. *et al.* Allosteric inhibition of complement function by a staphylococcal immune evasion protein. *Proceedings of the National Academy of Sciences* **107,** 17621–17626, doi: 10.1073/pnas.1003750107 (2010).
56. Frauenfelder, H., Petsko, G. A. & Tsernoglou, D. Temperature-Dependent X-Ray-Diffraction as a Probe of Protein Structural Dynamics. *Nature* **280,** 558–563, doi: 10.1038/280558a0 (1979).
57. Lipari, G. & Szabo, A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *Journal of the American Chemical Society* **104,** 4546–4559, doi: 10.1021/ja00381a009 (1982).
58. Lipari, G. & Szabo, A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results. *Journal of the American Chemical Society* **104,** 4559–4570, doi: 10.1021/ja00381a010 (1982).
59. Dong, M. D., Husale, S. & Sahin, O. Determination of protein structural flexibility by microsecond force spectroscopy. *Nature Nanotechnology* **4,** 514–517, doi: 10.1038/Nnano.2009.156 (2009).
60. Petoukhov, M. V. & Svergun, D. I. Applications of small-angle X-ray scattering to biomacromolecular solutions. *International Journal of Biochemistry & Cell Biology* **45,** 429–437, doi: 10.1016/j.biocel.2012.10.017 (2013).
61. Burger, V. M., Gurry, T. & Stultz, C. M. Intrinsically Disordered Proteins: Where Computation Meets Experiment. *Polymers* **6,** 2684–2719, doi: 10.3390/polym6102684 (2014).
62. Tallant, C. *et al.* Molecular Basis of Histone Tail Recognition by Human TIP5 PHD Finger and Bromodomain of the Chromatin Remodeling Complex NoRC. *Structure* **23,** 80–92, doi: http://dx.doi.org/10.1016/j.str.2014.10.017 (2015).
63. Mylonas, E. & Svergun, D. I. Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering. *J. Appl. Crystallogr.* **40,** s245–s249 (2007).

Acknowledgements

This was supported, in part, by a Steve G. and Renee Faculty Innovation Fellowship and an Early Postdoc.Mobility Fellowship from the Swiss National Science Foundation (SNSF) to V.M.B.

Author Contributions

D.J.A., V.M.B. and C.M.S. designed the experiments. D.J.A., V.M.B. and C.M.S. analyzed the data and wrote the paper.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Burger, V. M. *et al.* A Structure-free Method for Quantifying Conformational Flexibility in proteins. *Sci. Rep.* **6**, 29040; doi: 10.1038/srep29040 (2016).