Computing, analyzing and comparing the radius of gyration and hydrodynamic radius in conformational ensembles of intrinsically disordered proteins

Mustapha Carab Ahmed[1] · **Ramon Crehuet**[1,2] · **Kresten Lindorff-Larsen**[1]

August 15, 2019

Abstract The level of compaction of an intrinsically disordered protein may affect both its physical and biological properties, and can be probed via different types of biophysical experiments. Small-angle X-ray scattering (SAXS) probe the radius of gyration (R_g) whereas pulsed-field-gradient nuclear magnetic resonance (NMR) diffusion, fluorescence correlation spectroscopy and dynamic light scattering experiments can be used to determine the hydrodynamic radius (R_h). Here we show how to calculate R_g and R_h from a computationally-generated conformational ensemble of an intrinsically disordered protein. We further describe how to use a Bayesian/Maximum Entropy procedure to integrate data from SAXS and NMR diffusion experiments, so as to derive conformational ensembles in agreement with those experiments.

Keywords Radius of gyration, Hydrodynamic radius, Conformational ensemble, Compaction, Intrinsically disordered protein

1 Introduction

In contrast to natively folded proteins, intrinsically disordered proteins (IDPs) generally lack well-defined three-dimensional structures. Consequently, they explore a large number of distinct conformations, and their conformational properties are thus best described in statistical terms. One useful and informative way of representing this large conformational ensemble is through a distribution of the radius of gyration (R_g) of the IDP. The ensemble average $\langle R_g \rangle$ gives a rough measure of how compact a protein is and may, for example, be compared to the values for other proteins of similar lengths.

For a given configuration of a protein, the R_g may be calculated as the mass-weighted root mean distance to the centre of mass:

$$R_g = \left(\frac{\sum_i \|\mathbf{r}_i\|^2 m_i}{\sum_i m_i} \right)^{\frac{1}{2}} \tag{1}$$

where m_i is the mass of atom i and $\mathbf{r}_i$ is the position of atom i with respect to the center of mass of the molecule.

Experimentally, one may obtain an estimate of the ensemble-averaged value of the R_g of a protein by a Guinier analysis of small angle X-ray scattering (SAXS) profiles (***1***) or using various extended models of the scattering data (***2***, ***3***). For the sake of simplicity, we will loosely refer to the experimental value as R_g, omitting the bracket notation and only use brackets for explicitly

[1]Structural Biology and NMR Laboratory, Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen.
Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark
lindorff@bio.ku.dk
[2]Institute for Advanced Chemistry of Catalonia (IQAC-CSIC), c/ Jordi Girona 18-26, 08034 Barcelona, Spain

averaging computed values. Here, we note also that R_g calculated using Eq. 1 is not directly comparable to that obtained from analyses of SAXS data due to contributions to the scattering data from the solvent layer around the disordered protein (***4***, ***5***).

Similarly, but via different physical principles, the hydrodynamic radius of a protein also reports on the overall expansion of a protein. The hydrodynamic radius (R_h), also called the Stokes radius, is defined as the radius of a theoretical hard sphere that would have the same translational diffusion coefficient as the considered particle. The translational diffusion coefficient (D_t) of a protein may in turn be determined e.g. by pulsed-field gradient Nuclear Magnetic Resonance (NMR) diffusion experiments, fluorescence correlation spectroscopy and dynamic light scattering measurements, and is related to R_h through the Stokes-Einstein equation (***6***):

$$D_\mathrm{t} = \frac{k_\mathrm{B}T}{6\pi\eta R_\mathrm{h}} \tag{2}$$

where k_B is the Boltzmann constant, T is the temperature and η is the viscosity of the solvent.

Because both R_g and R_h probe the compaction of a disordered protein, and because they may contain complementary information about the distribution of states (***7***) there have been several studies on the relationship between the R_g and R_h for disordered proteins and polymers (***7***–***10***).

One such approach uses hydrodynamic modelling of protein conformations (***11***–***13***) to relate protein structure to R_h (***7***, ***10***). In line with theoretical expectations, the authors found that the ratio R_g/R_h depends substantially on the compaction of the protein chain, so that compact states have ratios ≈ 0.8 and expanded conformations have ratios between 1.2–1.6. Because the relative level of compaction of the chain, when quantified by R_g, also depends on the chain length, the ratio R_g/R_h also depends on the number of residues of the protein (N). Recently, these two effects were combined into a single, physically-motivated and empirically parameterized equation that enables one to calculate R_h for a configuration of an IDP from its R_g (***14***):

$$\frac{R_\mathrm{g}}{R_\mathrm{h}}(N, R_\mathrm{g}) = \frac{\alpha_1\left(R_\mathrm{g} - \alpha_2 N^{0.33}\right)}{N^{0.60} - N^{0.33}} + \alpha_3 \tag{3}$$

In addition to R_g and N (number of residues of the protein chain), the equation contains three parameters that were fitted to maximize agreement between the model and hydrodynamic calculations ($\alpha_1 = (0.216 \pm 0.001)$Å , $\alpha_2 = (4.06 \pm 0.02)$Å , and $\alpha_3 = (0.821 \pm 0.002)$Å). As discussed further below, since conformational averaging acts on the diffusion properties, the ensemble averaged value that should be compared to an experimentally measured R_h will not in general be the same as the linear average over the values of each conformation ($\langle R_h \rangle$). Also, note that the equation was parameterized using R_g values calculated from the C_α coordinates only. Values of R_g calculated in this way are generally very close to those calculated from all protein atoms, but this parameterization makes it possible to use the approach to calculate R_h also for coarse-grained C_α-only models.

Here we provide a step-by-step protocol to calculate R_g and subsequently R_h using Eq. 3 from a computationally generated conformational ensemble of an IDP. Together with calculations of SAXS data from simulations it is possible to compare the simulations to measurements of compaction. In cases where the computed and experimental quantities are not in perfect agreement, one may go one step further and refine the computational ensemble using the experimental data. We thus also demonstrate how to refine the ensembles by integrating experimental SAXS and R_h measurements, and thereby generate conformational ensembles that both take into account the physical principles encoded in the simulations as well as information from experiments. In addition to the motivation and description provided in this paper we also make available a Jupyter (Python) notebook with guided examples for performing analysis and generating many of the figures discussed here. We do, however, not provide instructions for how to generate conformational ensembles, and the reader is expected to have a basic understanding of the Python programming language to use the examples presented.

2 Materials

Experimental data and sequence of Sic1

- We used the following sequence for the Sic1: `GSMTPSTPPR SRGTRYLAQP SGNTSSSALM QGQKTPQKPS QNLVPVTPST TKSFKNAPLL APPNSNMGMT SPFNGLTSPQ RSPFPKSSVK RT`
- SAXS data for Sic1 (***15***) obtained from the Protein Ensemble Database (***16***) entry PED9AAA. (`http://pedb.vib.be/accession.php?ID=PED9AAA`
- We used the previously measured (***17***) experimental value of Rh (21.5 ± 1.1Å)

Software:

- Flexible Meccano (***18***) available from `http://www.ibs.fr/research/scientific-output/software/flexible-meccano/?lang=en`
- CAMPARI v3.0 (***19***) available from `https://sourceforge.net/projects/campari/`
- PULCHRA v3.06 (***20***) available from `http://www.pirx.com/pulchra/index.shtml`
- Pepsi-SAXS v1.4 (***21***) available from `https://team.inria.fr/nano-d/software/pepsi-saxs/`
- BME (***22***) available from `https://github.com/KULL-Centre/BME`
- MDtraj v1.9.3 (***23***) available from `http://mdtraj.org/1.9.3/`
- A Python Jupyter notebook (`https://jupyter.org/`) for performing the calculations and analyses described in this paper is available from `https://github.com/KULL-Centre/papers/edit/master/2019/IDP-methods-Ahmed-et-al/`

3 Methods

3.1 Generating Ensembles

We have chosen the 90 amino acid residues long protein Sic1 as an example for our calculations, as this protein has been studied extensively by both SAXS and various NMR methods (***15***, ***17***). We used Campari (***19***) and Flexible-Meccano (***18***) to generate two conformational ensembles of Sic1 in its unphosphorylated state. In the ensemble we generated using Campari (Ensemble 1) we used Monte Carlo sampling with the ABSINTH v3.2 implicit solvent model (***24***) and a temperature of 298K. The Sic1 protein was contained in a spherical simulation cell with a radius of 150 Å and an ion concentration of ≈140 mM, matching the experimental condition (***15***). For the Flexible-Meccano ensemble we generated conformations sampling random coil configurations as described (***18***). As Flexible-Meccano only generates a model of the protein backbone, we used PULCHRA (***20***) with default settings to add side chains to these structures and generate Ensemble 2. These side chain coordinates are necessary when we calculate SAXS data from the conformational ensembles. In total we generated 32,000 structures for Ensemble 1 and 10,000 structures for Ensemble 2.

3.2 Calculating Rg and Rh from ensembles

Many simulation and protein analysis software packages have the option of calculating the R_g of the protein. In this example we will use readily available and open source software. For calculating the R_g of the conformations we use MDTraj, a python module for protein analysis (***23***). Below we provide Python code demonstrating how to load the ensemble and calculate R_g for each structure, and then calculate R_h for each structure using Eq. 3. In the example we have collected all conformations of the ensemble in a trajectory file (here Ensemble1.trr). Depending on the file format of the trajectory file, one may also need a coordinate file (structure.pdb) or a topology file. Once these files are loaded, MDTraj is then used to calculate R_g for each structure in the ensemble, which in turn is converted into R_h using Eq. 3.

```
## Loading trajectory and calculate Rg
import mdtraj as md
traj = md.load("Ensemble1.trr",top="structure.pdb")

# Select CA atoms
CA_atoms = traj.topology.select('name CA')
traj.atom_slice(CA_atoms, inplace=True)

# Calculate Rg
rg = md.compute_rg(traj)
```

```
# Convert Rg from nm to Angstrom as equation
# from Nygaard et al uses Rg in Angstrom
Rg = rg*10

# Number of amino acids in protein
N = traj.n_residues

# Function for Rh calculation
def getRh(rg,N):
    # Parameters fitted in Nygaard et al.
    a1=0.216
    a2=4.06
    a3=0.821
    return (rg)/((a1*(rg-a2*N**(0.33)))/(N**(0.60)-N**(0.33))+a3)

# Rh calculation
Rh = getRh(Rg,N)
```

Once R_g and R_h have been calculated for each structure, these can be used to generate histograms of R_g (Fig. 1a) and R_h (Fig. 1b), and the average R_h can be calculated as for comparison to experimental values (see Note 1). We also show the calculated average of R_g in the plots (Fig. 1) (see Note 2) though as explained below, a better comparison to the experimental data requires calculations of SAXS intensities from the conformational ensemble (see also Note 3).

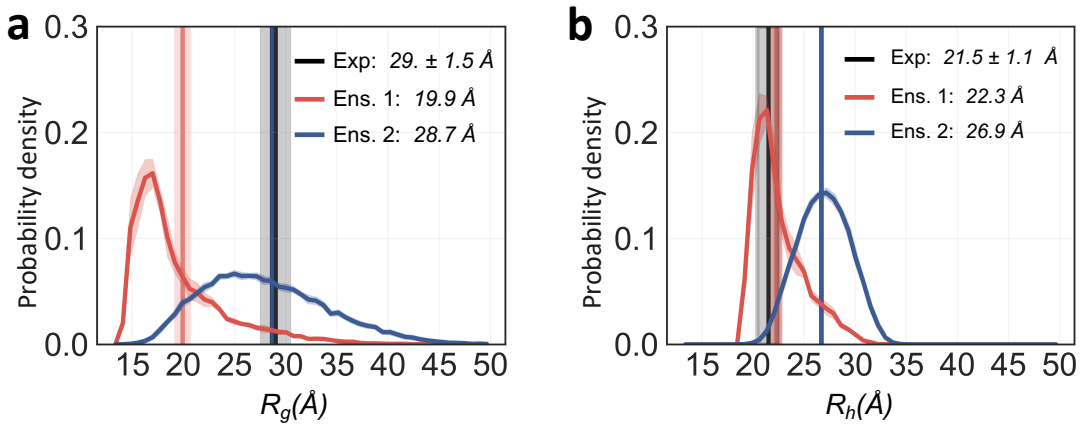

Fig. 1 Analyzing compaction in ensembles of Sic1. Probability distribution of **(a)** R_g and **(b)** R_h calculated from two ensembles that we generated of Sic1. Here, Ensemble 1 (red) was generated using Campari and Ensemble 2 (blue) was generated using Flexible-Meccano as described in the main text. **(a)** Solid vertical lines represents the ensemble average R_g ($\langle R_g \rangle_{trans}$; see Note 2 for the definition) of Ensemble 1 (red) and Ensemble 2 (blue). **(b)** Solid vertical lines represents the ensemble average R_h ($\langle R_h \rangle_{trans}$) calculated using Eq. 3 and as discussed in Note 1 from Ensemble 1 (red) and Ensemble 2(blue). The experimental values of R_g and R_h are shown in black. The error of the distribution and averages of R_g and R_h (shown as shaded areas) were estimated by block averaging using five blocks.

As described above the ratio R_g/R_h depends substantially on the compaction of the protein chain, so that compact states have ratios ≈ 0.8 and expanded conformations have ratios between 1.2–1.6 (***14***). For a protein of 91 amino acids, the switch-over point where the $R_g/R_h = 1$ lies at conformations with $R_g \approx 27$Å (see Note 4). Thus, conformations with $R_g < 27$Å have $R_h > R_g$ whereas conformations with $R_g > 27$Å have $R_h < R_g$. In this way, the distribution of R_h is 'pushed' towards the middle and has less density in the tails compared to the distribution of R_g (Fig. 1).

The distributions of R_g (Fig. 1a) and R_h (Fig. 1b) from Ensemble 1 and Ensemble 2 and the resulting averages, can also be compared to the experimental values from SAXS and NMR (***15***, ***17***). These results reveal two different scenarios for the two ensembles. First, $\langle R_h \rangle_{trans}$ calculated from Ensemble 1 (Campari) is in good agreement with the experimentally-determined value of R_h (Fig. 1b). At the same time, the calculated value of $\langle R_g \rangle_{trans}$ is substantially lower

than the average R_g value estimated from SAXS experiments (Fig. 1a). Second, for Ensemble 2 (Flexible-Meccano) we observe the opposite scenario, where the calculated $\langle R_g \rangle_{trans}$ is close to the value estimated by SAXS (Fig. 1a), and the calculated $\langle R_h \rangle_{trans}$ is substantially greater than the experimental value (Fig. 1b).

Disagreement between experiment and simulation is often indicative of problems with the molecular force fields or sampling (***25***), though differences may also arise from problems in e.g. the model used to calculate experimental data from structural ensembles (***5***, ***26***). While it is possible to improve molecular force fields directly against experimental data (***27***), we below describe how one can refine a specific ensemble against one or more sets of experimental measurements.

3.3 A Bayesian/Maximum Entropy approach

Above we have analysed two ensembles and used Eq. 3 to estimate R_h which in turn could be averaged and compared to NMR diffusion experiments. We also calculated R_g from the protein coordinates, though as noted this value is not directly comparable to the experimental measurements due to solvation effects (***5***). Nevertheless, the results suggested discrepancies between experiments and simulations.

Although there has been continued improvements in methods and force fields for sampling the conformational landscape of IDPs, it is still not uncommon that simulations are not in perfect agreement with experiments. In such cases, it is possible to bias the simulation to construct an ensemble that is in better agreement than the unbiased ensemble (***22***, ***28***–***31***).

We here use such a method to construct two new ensembles by reweighting the Campari and Flexible-Meccano ensembles with the experimental data, thus obtaining ensembles that are in better overall agreement with the SAXS and NMR diffusion experiments. Specifically, we use experimental SAXS data (***15***) and NMR diffusion measurements of R_h (***17***), and use our recently described Bayesian/Maximum Entropy (BME) protocol to reweight the conformational ensembles (***22***). We focus solely on the technical details of the approach rather than the biological relevance. Also, we exemplify using two experimental measures of compaction, but the approach is more generally applicable (See Note 5).

Briefly described, BME is based on a combined Bayesian/Maximum Entropy framework, and enables one to refine a simulation using multiple sources of (potentially noisy) data. The purpose of the reweigthing is to derive a new set of weights for each configuration in a previously generated ensemble so that the reweighted ensemble satisfies two criteria: (i) it matches the experimental data better than the original ensemble and (ii) it achieves this improved agreement by a minimal perturbation of the original ensemble. For additional details see Bottaro et al. (***22***) and references therein. In the current examples, both ensemble 1 and 2 were generated as unbiased ensemble and so the initial weights of all structures are uniform ($w_j^0 = 1/n$), where n is the number of structures in the ensemble.

The reweighting approach described above may in practice be achieved by updating the weights, w_j, of each configuration in the input ensemble by minimizing a function (the negative log-likelihood) (***22***, ***28***):

$$\mathcal{L}(w_1 \dots w_n) = \frac{1}{2}\chi^2(w_1 \dots w_n) - \theta S_{\text{rel}}(w_1 \dots w_n). \tag{4}$$

Here, the χ^2 quantifies the agreement between the experimental data and the corresponding values calculated from the reweighted ensemble. The second term contains the relative entropy, S_{rel}, which measures the deviation between the original ensemble and the reweighted ensemble $S_{\text{rel}} = -\sum_j^n w_j \log\left(\frac{w_j}{w_j^0}\right)$. The temperature-like parameter θ tunes the balance between fitting the data accurately (low χ^2) and not deviating too much from the prior (low S_{rel}). In practice, we determine this hyperparameter by evaluating the compromise between balancing the two terms in $\mathcal{L}$ (***22***, ***28***) (see also Note 6). When more than one set of experimental data is included in BME, the deviations between calculated and experimental values are summed in a global χ^2 function which is the sum of a χ^2 function for each set of data.

In practice it turns out that in many cases there is a more efficient approach to minimize $\mathcal{L}$ using the method of Lagrange multipliers, and this is the approach we take here (***22***, ***28***, ***32***) using the BME code, which is freely available at `https://github.com/KULL-Centre/BME`.

3.4 Calculating SAXS data from ensembles

The first step in the reweighting protocol is to collect the necessary data and structure it correctly for input in BME. We first calculate the SAXS intensity profiles by fitting to the experimental curve for each structure of the two ensembles using Pepsi-SAXS (***21***). Pepsi-SAXS has free parameters for the solvation layer that are calculated for each fit. To decrease the risk of overfitting, we used a two-step procedure. First, we fitted the parameters to each structure. Second, we calculate the averages of the resulting fitted values of the solvation parameters and re-ran Pepsi-SAXS with these parameters fixed to those averages. Alternative methods for calculating SAXS from conformational ensembles exist (***4***) and may also be used (See Note 7 and Note 8).

We then structure the input files as shown below for SAXS BME input. The experimental SAXS input file is structured such that it contains the following three columns: the momentum transfer (q), intensity ($I(q)$), and the error ($\sigma_I(q)$) (as shown below). Each of these three columns are m rows long, where m is the number of experimental data points. The input file for the calculated values contains n rows(number of structure in the ensemble), and $m+1$ columns. The first column is for labeling the individual structure/frame from the ensemble. Further details for how to structure the input files for other data can be found in the original description of BME and in the online examples (***22***).

Experimental file format:

$$
\begin{array}{l}
\texttt{\# DATA=SAXS PRIOR=GAUSS} \\
q_1 \quad I(q_1) \quad \sigma_1 \\
q_2 \quad I(q_2) \quad \sigma_2 \\
\vdots \quad \vdots \quad \vdots \\
q_{179} \quad I(q_{179}) \quad \sigma_{179}
\end{array}
$$

Simulation SAXS file format:

$$
\begin{array}{lccc}
\texttt{\# label} & q_1 & \cdots & q_{179} \\
\texttt{frame1} & I(q_1)_1^{\mathrm{CALC}} & \cdots & I(q_{179})_1^{\mathrm{CALC}} \\
\texttt{frame2} & I(q_1)_2^{\mathrm{CALC}} & \cdots & I(q_{179})_2^{\mathrm{CALC}} \\
\vdots & \vdots & & \vdots \\
\texttt{frame(n)} & I(q_1)_n^{\mathrm{CALC}} & \cdots & I(q_{179})_n^{\mathrm{CALC}}
\end{array}
$$

Once these calculations have been done, we may load the data in python and run BME:

```
## Import module
import bme_reweight as bme

# Initialize reweighting class
rew = bme.Reweight()

# Locate input files
exp_saxs = 'exp_saxs.txt'
calc_saxs = 'calc_saxs.txt'

# Load data to BME
rew.load(exp_saxs,calc_saxs)

# Optimize using theta=5
chi2_before,chi2_after, srel = rew.optimize(theta=5)

# Get updated weights
rewighted_weights=rew.get_weights()
```

3.5 Reweighting Sic1 ensembles against SAXS and NMR diffusion experiments

We used the methods described above to determine a reweighted ensemble of Sic1 that takes into account both the prior information encoded in the initial ensemble (from Campari or Flexible-Meccano) as well as the experimental measurements of compaction from NMR diffusion and SAXS.

Before reweighting was applied, Ensemble 1 appears too compact when judged by agreement with the R_g-value extracted from the SAXS data, but is in good agreement with the NMR diffusion data (Fig. 1). In contrast, Ensemble 2 is in good agreement with the SAXS-derived R_g, but appears too expanded when compared to the NMR diffusion measurements (Fig. 1). The goal was therefore to examine whether one could construct an ensemble that provides a useful compromise between the two data sets. We note here that the NMR diffusion data were recorded at 278 K (***17***), whereas the SAXS data were obtained at room temperature (***15***), though we only expect a modest change in compaction in this temperature range (***33***, ***34***). We note also that our goal is not to discuss in detail the conformational ensemble of Sic1, but rather to showcase how one may combine different measures of compaction.

We reweighted the two ensembles against the NMR and SAXS data and compared to the unweighted ensembles (Fig. 2). The first step is to chose the temperature-like hyperparameter, θ, that sets the balance between fitting the data and not deviating too much from the input ensemble. The latter may be quantified by calculating the fraction of the frames in the input ensemble, $N_{eff} = \exp(S_{rel})$, that effectively contribute to the calculated ensemble averages after reweighting. Thus, $N_{eff} = 1$ corresponds to the initial unweighted ensemble and a low value of N_{eff} indicates that only a small fraction of the original ensemble has been selected to improve agreement with experiments. We scanned values of θ and calculated the agreement with both the SAXS and NMR diffusion data at each value of θ and for each of the two ensembles (Figs. 2a and 2b). Note that we here plot a reduced χ^2 (χ^2_{red}) for each of the two experiments individually , but that the optimization acts to reduce the sum of the two non-reduced χ^2-values. Since there is a 179 points in the SAXS measurements, this sum contains a large contribution from the SAXS data (see Note 8). In our analyses here, we chose $\theta = 100$ for Ensemble 1 and $\theta = 7$ for Ensemble 2, though in practical applications it would be advised to examine the results of other choices (See Note 6).

The effect of reweighting can be seen both on the distribution of R_g (Fig. 2c and 2d) and R_h (Fig. 2e and 2f). The more compact Ensemble 1 is shifted to include more expanded structures, bringing $\langle R_g \rangle_{trans}$ substantially closer to the value estimated from the SAXS data, while only increasing the calculated R_h value $\approx 15\%$ above the experimental value. Similarly, the more expanded Ensemble 2 is shifted to give greater weight to more compact configurations, bringing the calculated R_h closer to experiment while only shifting the $\langle R_g \rangle_{trans}$ down by $\approx 13\%$.

While it is convenient to examine the distribution of R_g before and after reweighting, the actual reweighting is done against the SAXS data not the estimated R_g. As explained above, the solvent layer around the protein also contributes to the SAXS measurements, and there may be ≈5–10% difference in the R_g calculated from the protein coordinates and the value estimated by SAXS (***5***). We thus also show the agreement between the experimental and calculated SAXS curves (Fig. 2g and 2h). It is clear that the reweighted SAXS curves are substantially closer to the experimental data, though there still remains some discrepancy in the low-q range for Ensemble 1.

3.6 Summary

We have shown here how it is possible to calculate R_h from a conformational ensemble using Eq. 3 and compare to experimental data obtained e.g. from NMR diffusion measurements. Such measurements provide an alternative view of the compaction to that obtained e.g. from SAXS experiments, and indeed it has previously been shown that simultaneous refinement against R_h and R_g can provide insight into the shape of the distribution of R_g (***7***).

We chose the protein Sic1 for exemplifying our analyses since the level of expansion has been measured for this protein using both SAXS and NMR diffusion measurements. Since the data were recorded at slightly different conditions and temperatures, we do not aim to make strong

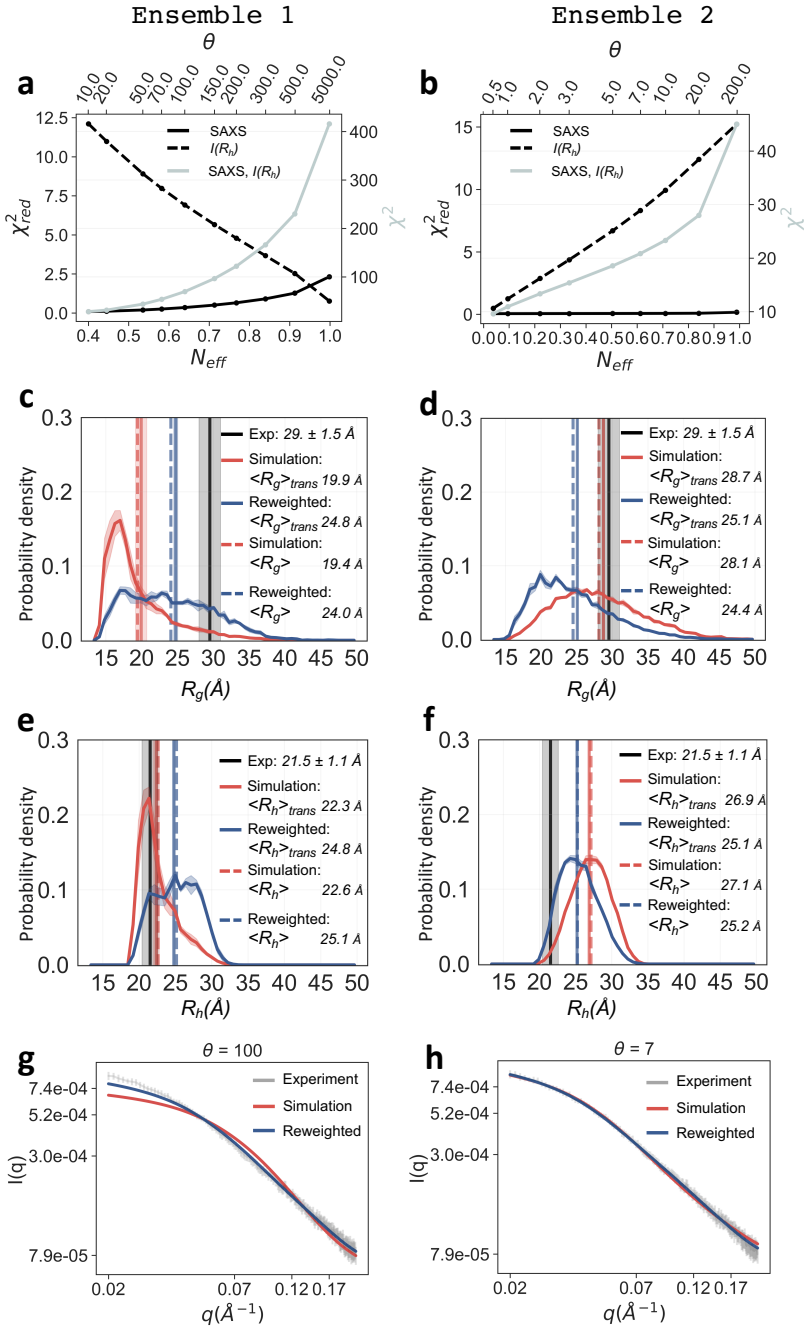

Fig. 2 Constructing ensembles to improve agreement with experiments. We used BME reweighting with SAXS and R_h data for Ensemble 1 (**a, c, e, g**) and Ensemble 2 (**b, d, f, h**). We label the R_h data as $I(R_h)$ as we here use intensity-based averaging of the measurements (Note 1). (**a,b**) We plot N_{eff} (the effective number of frames left after reweighting) vs. χ^2 when the scaling parameter θ is varied (top axis). The left axes show χ^2_{red} for each of the two experiments, whereas the right axis shows the total χ^2 that is the sum of the two (non-reduced) χ^2 values (Note 8). For further analyses we chose $\theta = 100$ (Ensemble 1) and $\theta = 7$ (Ensemble 2). (**c-f**) We show the distribution of R_g (**c,d**) and R_h (**e,f**) before (red) and after (blue) reweighting the two ensembles. Averages over these distributions (both before and after reweighting) are shown either as standard (linear) averages (dashed lines) or 'transformed' averages ($\langle R_g \rangle_{trans}$ and $\langle R_h \rangle_{trans}$ as described in Notes 1 and 2). (**g,h**) We show the calculated SAXS intensity from the original ensemble and the refined ensembles and compared to the experimental data. In panels c–f the experimental data are shown in black lines and the errors are shown as shades. Errors in calculated values were estimated by block averaging using 5 blocks.

conclusions about the conformational ensemble of Sic1, and have used it here mostly to showcase the methods for analyses.

We generated two ensembles and show that one is in relatively good agreement with the NMR diffusion data whereas the other is in better agreement with the SAXS data. At this moment the origins of these differences are unclear. Variation in experimental conditions such as temperature may affect both R_g and R_h (***34***). Also, it is possible that our approach for calculating R_h is not always sufficiently accurate since it is inherently limited to the accuracy achievable by hydrodynamic modelling (***14***), and an important question for future research is whether we can provide better models to link conformation and calculated values of R_h. Finally, despite continued improvement in methods for calculating SAXS data from ensembles (***4***) there are still potential sources of error from e.g. solvation effects (***5***). Nevertheless, we note that by reweighting the ensembles against both sets of experiments it is possible to construct an ensemble that provides a reasonable balance between the two. As more proteins are studied by both NMR and SAXS it should be possible to test and improve our relationship between R_g and R_h, thus enabling further insight into the rules that govern compaction of IDPs.

4 Notes

1. When calculating averages over ensembles, in particular for broad ensembles such as for IDPs, it is important to take the correct form of averaging into account. The best way to calculate averages over experimental quantities will depend both on the type of experiment and often also e.g. on the time scales for conformational averaging. Throughout this paper we make the assumption that averages can be calculated as time-independent averages over the conformational ensemble. In the case of measurements of the hydrodynamic radius, R_h, we have explored two different types of averaging. In case the experiment measures the average diffusion coefficient, then according to Eq. 2 then the average should be calculated as $\langle R_h \rangle_{trans} = \langle R_h^{-1} \rangle^{-1}$. Here we have introduced the notation $\langle R_h \rangle_{trans}$ to represent that the averaging takes place on a *transformed* value (in this case proportional to R_h^{-1}). When R_h is measured by pulsed-field gradient NMR diffusion measurements (***35***) the NMR signal intensity, I, is proportional to $\exp\left(-R_h^{-1}\right)$ and it may therefore be more appropriate to use this function to perform the averaging. In this case
$$\langle R_h \rangle_{trans} = -\ln\left(\langle \exp\left(-R_h^{-1}\right)\rangle\right)^{-1}$$
It is this intensity-based averaging that we use here, though in practice we have found it to give essentially the same result as using $\langle R_h^{-1} \rangle^{-1}$.
2. Similar to the issue of averaging R_h discussed in Note 1 above, we use $\langle R_g \rangle_{trans} = \langle R_g^2 \rangle^{1/2}$ when calculating averages over the radius of gyration. This kind of averaging mimics the averaging in the low-q range of SAXS curves. Note, however, that $\langle R_g \rangle_{trans}$ calculated in this way should not directly be compared to experimental values of R_g since the latter includes solvation effects.
3. Notes 1 and 2 discuss the transformations that are relevant for comparing calculated and experimental quantities. We note, however, that during the reweighting protocol and generally when one makes quantitative comparisons between experiments and computation it is in general better to compare to the direct experimental quantities. In the case of SAXS experiments we thus judge agreement and perform reweighting against the experimentally measured intensities. In the case of the R_h measured for Sic1 by NMR diffusion experiments, we transform the experimental value of R_h (and its error), as well as the values calculated for each structure using the function $I \propto \exp\left(-R_h^{-1}\right)$, as described in our associated Jupyter notebook. We note that in the future it might be more appropriate to perform such fitting to the measured intensities as a function of the gradient strength.
4. The level of compaction as quantified by the value of R_g at which $R_g = R_h$ (R_g^0) can be estimated by rearranging Eq. 3 to obtain: $R_g^0 = \alpha_1^{-1}(1-\alpha_3)(N^{0.60} - N^{0.33}) + (\alpha_2 - N^{0.33})$. For a protein with $N = 91$ one obtains $R_g^0 = 27$Å.
5. We have here described approaches to refine ensembles against SAXS and NMR diffusion measurements. The BME method has also been used for IDPs with NMR chemical shifts (***36***), and may also readily be applied to SANS data, NOEs, scalar couplings or other measuremenets that can be calculated as averages over configurational ensembles.

6. Currently, the value of the hyperparameter θ (which sets the balance between information from the data and the force field) is set manually. In certain cases it may be possible to set it via a cross-validation approach (***36***) or may be integrated out as a Bayesian 'nuissance parameter' (***28***)
7. We have here used Pepsi-SAXS to calculate X-ray scattering curves from a conformational ensemble due to its ease of use and the relatively high computational efficiency. The latter is particularly important for large conformational ensembles. We note, however, that several other methods exist and suggest users in particular to keep solvent effects in mind when calculating and interpreting SAXS data (***4***, ***5***). In the Jupyter notebook avialable online we provide a script that performs a two-pass run of Pepsi-SAXS to find a reasonable value of solvent-related parameters in the calculations.
8. When plotting χ^2_{red} in Fig. 2, we calculate it by normalizing χ^2 by the number of experimental data points: $\chi^2_{red} = m^{-1}\chi^2$. We note that this an approximation because the number of degrees of freedom can be smaller because different parameters are fitted such as parameters involved in calculated the SAXS curves. Also, in the case of reweighting the weights themselves may be considered as free parameters. Thus, we note that the reweighting does not involve this normalization, and that the χ^2_{red} is only shown in Fig. 2 to give the reader an impression of the level of agreement. We also note that when fitting the R_h the resulting sum in χ^2 only contains a single term. Finally, we note that we here simply combine the χ^2 from the SAXS and NMR diffusion experiments by adding up the two individual χ^2 terms. In the current implementation, BME does not enable automatic balancing of independent experiments and instead sets this balance by the error estimates of the individual experiments. We note, however, that while the SAXS data for Sic1 contains 179 individual data points, the amount of information in a SAXS experiments typically corresponds to a smaller number of parameters (***37***) and a more careful balance between the information in the SAXS and NMR diffusion experiments should take such effects into account (***38***).

5 Acknowledgements

We thank Dr. Tanja Mittag for providing feedback on the manuscript, Dr. Andreas Haahr Larsen for general discussions about SAXS experiments and calculations, and Dr. Martin Blackledge for suggesting to use intensity-based averaging for R_h. The research described here was supported by a grant from the Lundbeck Foundation to the BRAINSTRUC structural biology initiative.

References

1. Guinier A, Fournet G (1955) Small angle X-ray scattering. John Wiley and Sons,New York
2. Zheng W, Best RB (2018) An extended guinier analysis for intrinsically disordered proteins. Journal of molecular biology 430(16):2540–2553
3. Riback JA, Bowman MA, Zmyslowski AM, Knoverek CR, Jumper JM, Hinshaw JR, Kaye EB, Freed KF, Clark PL, Sosnick TR (2017) Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. Science 358(6360):238–241
4. Hub JS (2018) Interpreting solution x-ray scattering data using molecular simulations. Current opinion in structural biology 49:18–26
5. Henriques J, Arleth L, Lindorff-Larsen K, Skepö M (2018) On the calculation of saxs profiles of folded and intrinsically disordered proteins from computer simulations. Journal of molecular biology 430(16):2521–2539
6. Edward JT (1970) Molecular volumes and the stokes-einstein equation. Journal of Chemical Education 47(4):261
7. Choy WY, Mulder FA, Crowhurst KA, Muhandiram D, Millett IS, Doniach S, Forman-Kay JD, Kay LE (2002) Distribution of molecular size within an unfolded state ensemble using small-angle x-ray scattering and pulse field gradient nmr techniques. Journal of molecular biology 316(1):101–112
8. Burchard W, Schmidt M, Stockmayer W (1980) Information on polydispersity and branching from combined quasi-elastic and intergrated scattering. Macromolecules 13(5):1265–1272

9. Oono Y, Kohmoto M (1983) Renormalization group theory of transport properties of polymer solutions. i. dilute solutions. The Journal of Chemical Physics 78(1):520–528
10. Lindorff-Larsen K, Kristjansdottir S, Teilum K, Fieber W, Dobson CM, Poulsen FM, Vendruscolo M (2004) Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein. Journal of the American Chemical Society 126(10):3291–3299
11. de la Torre JG, Huertas ML, Carrasco B (2000) Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. Biophysical journal 78(2):719–730
12. Ortega A, Amorós D, De La Torre JG (2011) Prediction of hydrodynamic and other solution properties of rigid proteins from atomic-and residue-level models. Biophysical journal 101(4):892–898
13. Amorós D, Ortega A, García de la Torre J (2013) Prediction of hydrodynamic and other solution properties of partially disordered proteins with a simple, coarse-grained model. Journal of chemical theory and computation 9(3):1678–1685
14. Nygaard M, Kragelund BB, Papaleo E, Lindorff-Larsen K (2017) An efficient method for estimating the hydrodynamic radius of disordered protein conformations. Biophysical journal 113(3):550–557
15. Mittag T, Marsh J, Grishaev A, Orlicky S, Lin H, Sicheri F, Tyers M, Forman-Kay JD (2010) Structure/function implications in a dynamic complex of the intrinsically disordered sic1 with the cdc4 subunit of an scf ubiquitin ligase. Structure 18(4):494–506
16. Varadi M, Kosol S, Lebrun P, Valentini E, Blackledge M, Dunker AK, Felli IC, Forman-Kay JD, Kriwacki RW, Pierattelli R, et al (2013) pe-db: a database of structural ensembles of intrinsically disordered and of unfolded proteins. Nucleic acids research 42(D1):D326–D335
17. Mittag T, Orlicky S, Choy WY, Tang X, Lin H, Sicheri F, Kay LE, Tyers M, Forman-Kay JD (2008) Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. Proceedings of the National Academy of Sciences of the United States of America 105(46):17,772–17,777
18. Ozenne V, Bauer F, Salmon L, Huang Jr, Jensen MR, Segard S, Bernadó P, Charavay C, Blackledge M (2012) Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics 28(11):1463–1470
19. Vitalis A, Pappu RV (2009) Methods for monte carlo simulations of biomacromolecules. Annual reports in computational chemistry 5:49–76
20. Rotkiewicz P, Skolnick J (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. Journal of computational chemistry 29(9):1460–1465
21. Grudinin S, Garkavenko M, Kazennov A (2017) Pepsi-SAXS: An adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. Acta Crystallogr D 73:449–464
22. Bottaro S, Bengtsen T, Lindorff-Larsen K (2018) Integrating molecular simulation and experimental data: A bayesian/maximum entropy reweighting approach. bioRxiv p 457952
23. McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, Schwantes CR, Wang LP, Lane TJ, Pande VS (2015) Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. Biophysical Journal 109(8):1528 – 1532
24. Vitalis A, Pappu RV (2009) Absinth: A new continuum solvation model for simulations of polypeptides in aqueous solutions. Journal of computational chemistry 30(5):673–699
25. Bottaro S, Lindorff-Larsen K (2018) Biophysical experiments and biomolecular simulations: A perfect match? Science 361(6400):355–360
26. van Gunsteren WF, Daura X, Hansen N, Mark AE, Oostenbrink C, Riniker S, Smith LJ (2018) Validation of molecular simulation: an overview of issues. Angewandte Chemie International Edition 57(4):884–902
27. Norgaard AB, Ferkinghoff-Borg J, Lindorff-Larsen K (2008) Experimental parameterization of an energy function for the simulation of unfolded proteins. Biophysical journal 94(1):182–192
28. Hummer G, Köfinger J (2015) Bayesian ensemble refinement by replica simulations and reweighting. The Journal of chemical physics 143(24):12B634_1
29. Boomsma W, Ferkinghoff-Borg J, Lindorff-Larsen K (2014) Combining experiments and simulations using the maximum entropy principle. PLoS Comput Biol 10(2):e1003,406

30. Bonomi M, Heller GT, Camilloni C, Vendruscolo M (2017) Principles of protein structural ensemble determination. Curr Opin Struct Biol 42:106–116
31. Cesari A, Reißer S, Bussi G (2018) Using the maximum entropy principle to combine simulations and solution experiments. Computation 6(1):15
32. Cesari A, Gil-Ley A, Bussi G (2016) Combining simulations and solution experiments as a paradigm for rna force field refinement. Journal of Chemical Theory and Computation 12(12):6192–6200
33. Kjaergaard M, Nørholm AB, Hendus-Altenburger R, Pedersen SF, Poulsen FM, Kragelund BB (2010) Temperature-dependent structural changes in intrinsically disordered proteins: Formation of α–helices or loss of polyproline ii? Protein Science 19(8):1555–1564
34. Jephthah S, Staby L, Kragelund BB, Skep M (2019) Temperature dependence of intrinsically disordered proteins in simulations: What are we missing? Journal of Chemical Theory and Computation 15(4):2672–2683
35. Wilkins DK, Grimshaw SB, Receveur V, Dobson CM, Jones JA, Smith LJ (1999) Hydrodynamic radii of native and denatured proteins measured by pulse field gradient nmr techniques. Biochemistry 38(50):16,424–16,431
36. Crehuet R, Jorro PJB, Lindorff-Larsen K, Salvatella X (2019) Bayesian-maximum-entropy reweighting of idps ensembles based on nmr chemical shifts. BioRxiv p 689083
37. Vestergaard B, Hansen S (2006) Application of bayesian analysis to indirect fourier transformation in small-angle scattering. Journal of applied crystallography 39(6):797–804
38. Larsen AH, Arleth L, Hansen S (2018) Analysis of small-angle scattering data using model fitting and bayesian regularization. J Appl Crystallogr 51(4):1151–1161