

SAS profile correlations reveal SAS hierarchical nature and information content

Yannick G. Spill, Michael Nilges

▶ **To cite this version:**

Yannick G. Spill, Michael Nilges. SAS profile correlations reveal SAS hierarchical nature and information content. PLoS ONE, 2017, 12 (5), pp.e0177309. 10.1371/journal.pone.0177309 . pasteur-01570231

HAL Id: pasteur-01570231

https://pasteur.hal.science/pasteur-01570231

Submitted on 28 Jul 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

PLOS ONE

RESEARCH ARTICLE

SAS profile correlations reveal SAS hierarchical nature and information content

Yannick G. Spill¤*, Michael Nilges

Structural Bioinformatics Unit, Department of Structural Biology and Chemistry, Institut Pasteur, 25 rue du Docteur Roux, 75015 Paris, France

¤ Current address: CNAG-CRG, Torre I planta 10, Parc Cientific de Barcelona, Carrer de Baldiri Reixac 4, 08028 Barcelona, Spain
* yannick.spill@nsup.org

Abstract

In structural biology, Small-Angle Scattering experiments (SAS) are unique, because although they provide low resolution data, they can be performed in closer-to-native conditions than those arising in X-Ray crystallography. A number of questions on SAS, however, remain unsolved, particularly in the light of modelling ensembles of conformers in solution. In this article, we study the ensemble average and covariance of SAS profiles analytically. Using this ensemble covariance, we demonstrate the hierarchical nature of SAS profiles. Furthermore, we show that the information content is not uniform and reaches its maximum in the intermediate *q* range. The arguments are generalized using microsecond-scale molecular dynamics trajectories of the lysozyme and on an ensemble of the intrinsically disordered protein p15PAF. We show that for highly flexible systems, the SAS profile is a representation of the ensemble of conformers in solution, and not that of one conformer in particular.

Check for updates

OPEN ACCESS

Citation: Spill YG, Nilges M (2017) SAS profile correlations reveal SAS hierarchical nature and information content. PLoS ONE 12(5): e0177309. https://doi.org/10.1371/journal.pone.0177309

Editor: Bruce R. Donald, Duke University, UNITED STATES

Received: November 2, 2016

Accepted: April 25, 2017

Published: May 11, 2017

Copyright: © 2017 Spill, Nilges. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was funded by the European Research Commission (Advanced Grant ERC-2011-StG 294809 BayCellS, to MN).

Competing interests: The authors have declared that no competing interests exist.

Introduction

Biological small-angle scattering (SAS) of X-rays (SAXS) or neutrons (SANS) has regained interest, judging by the technical improvements made to beamlines recently [1, 2]. SAS experiments are easier to perform and in closer-to-native conditions than X-ray crystallography. Therefore, SAS is in a unique position for structural biologists, and the generalization of in-house SAXS experiments will only strengthen this position.

However convenient SAS experiments are, they only provide a limited amount of information. They are therefore often combined with other experiments to reach atomic resolution [3]. It is frequent to extract a number of parameters, such as the radius of gyration, the Porod exponent, or the volume of correlation [4–8]. Simple parameters, such as thos extracted from Kratky or Porod-Debye plots, can be used to assess the flexibility of a macromolecule [9]. Whether these or other parameters are independent of each other, and how they relate to the maximum number of independent points in a SAS profile [10–12] has not been studied in-depth.

It is also becoming clear that SAS measures conformational diversity [13]. More than 60% of all articles on the topic of SAXS ensembles have been published in the last five years, following the development of a number of methods for ensemble modelling (EOM [14], MES [15], BSS-SAXS [16], EROS [17], SES [18] and BE-SAXS [19]. In these methods, the SAS profile is almost always modelled as a weighted average of the profiles of the individual conformations. The different methods differ by the way they select the weights, and the number of conformations.

These methods are best suited to describe a small number of well-defined conformations present simultaneously in solution. However, cases where conformations vary continuously from one to the other are to be treated with much more care. As noted early on [14, 15], the obtained ensemble is then illustrative of the diversity of possible conformations. The number of conformations these methods propose are then not necessarily to be taken as granted, because the conformations are expected to have a strong internal variability [20]. In that respect, EROS [17] goes further in modelling continuous motion, because each conformation is already an average over a potentially large number of structures. Yet, the number of parameters, which in essence is three times the number of atoms times the number of structures, still becomes very large for such systems, and the risk of overfitting is not negligible. The most promising approach in that respect is the recently proposed BE-SAXS [19]. It proposes a generative model for the protein ensemble fitted on experimental SAXS data. This model therefore controls the expansion of the number of parameters. Yet it is unclear how that number of parameters can be extracted from it, and how to summarize the obtained distribution. Clearly, additional ways to represent continuous conformational variability would be welcome in the field. A first step is therefore to describe how structural variability affects SAS profiles in solution, which is the aim of this article.

Materials and methods

We used the two 1 μs Molecular Dynamics (MD) simulations of the lysozyme described by Po-chia Chen and Jochen S. Hub [21], dropping the first 100 ns in each simulation. We performed the most likely alignment of the remaining frames using THESEUS [22]. This alignment produced a clash-free median structure for which the median atomic fluctuation was $\tau = 0.5$ Å. It corresponds to the structure in the input which is closest to the center of the cluster. The median structure of the first simulation was taken as the center structure for all analytical calculations (and Figs 1 and 2).

Because we do not want to discuss the impact of solvation models on the calculations, we used a single model (FoXS [23, 24] with $c_1 = 1$ and $c_2 = 0$) to compute the SAS profiles of all structures of the simulations; more accurate solvation models should however be employed for practical applications when long MD trajectories are available [21, 25–27]. We refer to this as the correlated dataset. The numerical SAS variance of the lysozyme was then obtained by computing the variance matrix of these SAS profiles. We do not expect other solvation models to be very different from the two cases presented here.

Extension to intrinsically disordered proteins was performed on the p15PAF ensemble [28], available in the protein ensemble database under the accession code PED6AAA. We used the experimental profile of p15PAF, and the 4939 structures comprised in the ensemble. Individual SAS profiles were calculated with FoXS using $c_1 = 1$ and $c_2 = 0.63$.

Results

Ensemble average and covariance of SAS profiles

In this article, we relate the SAS profiles of conformers arising naturally in solution through thermal motion. We start with the Debye formula of the scattering intensity at momentum

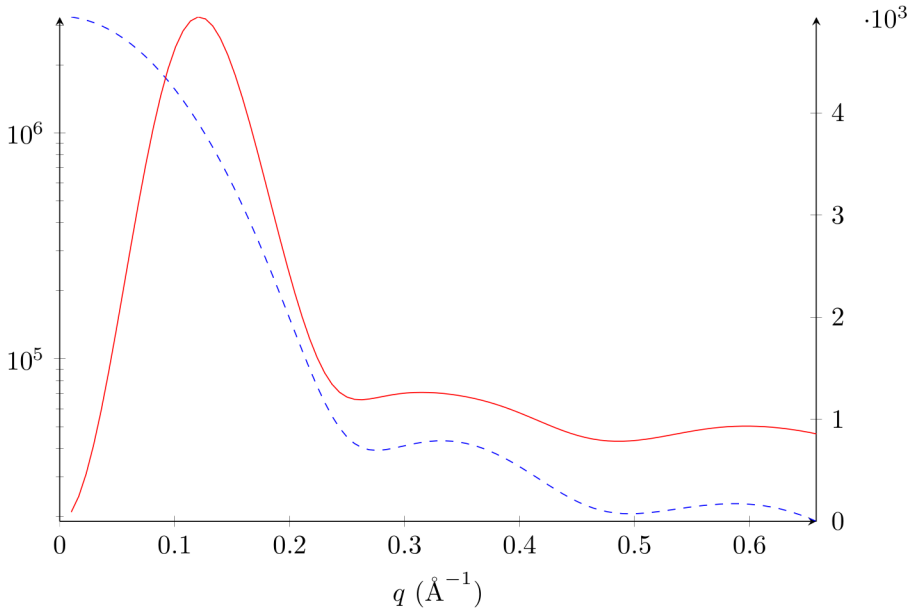


Fig 1. Lysozyme average SAS profile and standard deviation. Average SAS profile in blue dashed line, left axis, in arbitrary units, Eq 6. Standard deviation in solid red line, right axis, in arbitrary units, square root of Eq 11 with $q_i = q_j$. All calculations use $\tau = 0.5$ Å and FoXS form factors with $c1 = 1$ and $c2 = 0$ [23, 24].

https://doi.org/10.1371/journal.pone.0177309.g001

transfer $q \equiv 4\pi \sin\theta/\lambda$ (scattering angle 2θ and wavelength λ) for an atomic structure comprising N atoms whose coordinates define the vector X

$$I_X(q) = \sum_{k=1}^{N}\sum_{l=1}^{N} f_k(q) f_l(q) \frac{\sin(qd_{kl})}{qd_{kl}} \tag{1}$$

where d_{kl} is the Euclidean distance between atoms k and l and $f_k(q)$ is the form factor of atom k at q [29]. Form factors used in this formula must include volume exclusion and solvent effects. Their definition is not a trivial task and falls outside of the scope of this article.

We now treat X as a random vector having $3N$ components. To model thermal motion we assume that X follows a Normal distribution around a mean structure x° with a diagonal covariance matrix such that atom k has variance τ_k^2 along each of its coordinates. This simplifying assumption allows us to obtain analytical formulæ; it is the same as that used for the Debye-Waller temperature factors [30, 31]. We discuss generalizations thereof further down.

Average intensity

The average intensity is computed by taking the mathematical expectation $\mathbb{E}(I_X(q))$ of the intensity $I_X(q)$ over X. Using the linearity of the expectation in the Debye formula (Eq 1), we have

$$\mathbb{E}(I_X(q)) = \sum_k f_k(q)^2 + \sum_{k\neq l} f_k(q) f_l(q) \mathbb{E}\left(\frac{\sin(qd_{kl})}{qd_{kl}}\right) \tag{2}$$

Therefore, we seek the average of $\frac{\sin(qd_{kl})}{qd_{kl}}$ for any pair of atoms k and l. It can be shown that in this case, the distance d between these two atoms follows a noncentral χ distribution with

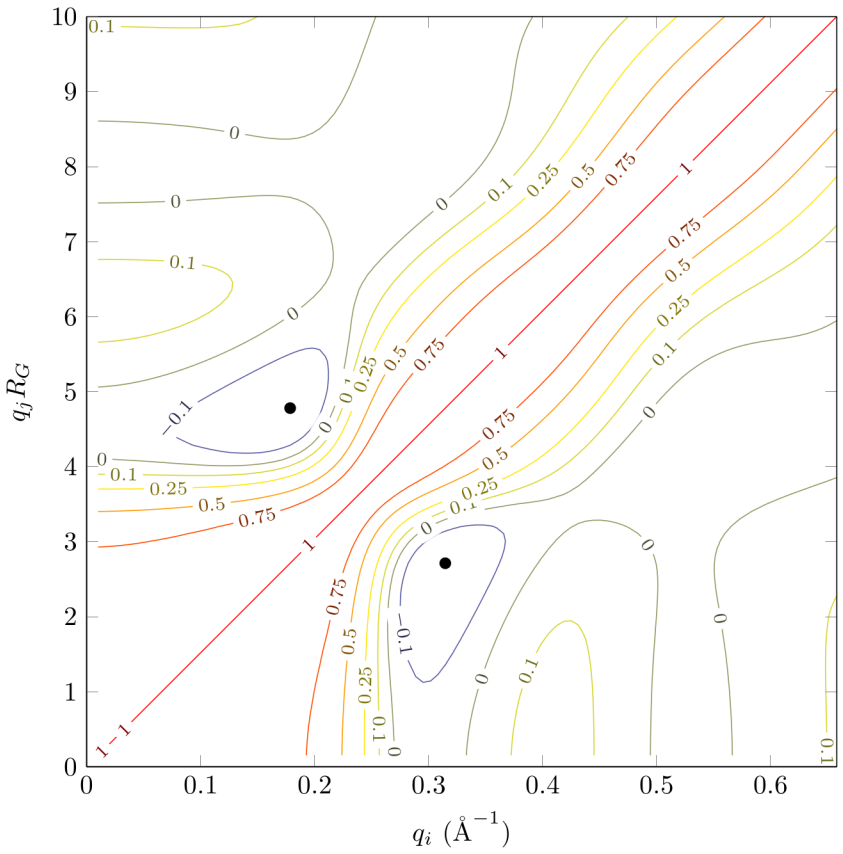


Fig 2. Contour plot of lysozyme SAS profile correlations. The correlations are given by $\rho(q_i, q_j)$ (Eq 10). R_G = 15.2 Å. Smallest correlation is -0.28 and is indicated by a blue dot. Calculations use τ = 0.5 Å and FoXS form factors with $c1$ = 1 and $c2$ = 0 [23, 24].

https://doi.org/10.1371/journal.pone.0177309.g002

three degrees of freedom, whose probability density function is

$$p_\chi(d_{kl}|d^\circ_{kl},\tau_k,\tau_l) = \frac{1}{\sqrt{2\pi(\tau_k^2+\tau_l^2)}}\frac{d_{kl}}{d^\circ_{kl}}\left[\exp\left(-\frac{(d_{kl}-d^\circ_{kl})^2}{2(\tau_k^2+\tau_l^2)}\right) - \exp\left(-\frac{(d_{kl}+d^\circ_{kl})^2}{2(\tau_k^2+\tau_l^2)}\right)\right] \quad (3)$$

where d°_{kl} is the distance obtained when the atoms are at their average positions (see S1 Text). Without any approximation, we thus have

$$\forall q \geq 0, \quad \mathbb{E}\left(\frac{\sin(qd_{kl})}{qd_{kl}}\right) = \frac{\sin(qd^\circ_{kl})}{qd^\circ_{kl}}\exp\left(-q^2\frac{\tau_k^2+\tau_l^2}{2}\right) \quad (4)$$

This equality can then be inserted in the Debye equation to yield the SAS profile of the ensemble of conformations centered at structure x° described by the normal random variable X, now referred to as *thermal ensemble*.

$$\mathbb{E}(I_X(q)) = \sum_{k=1}^{N}\sum_{l=1}^{N}\left(f_k(q)e^{-q^2\tau_k^2/2}\right)\left(f_l(q)e^{-q^2\tau_l^2/2}\right)\frac{\sin(qd^\circ_{kl})}{qd^\circ_{kl}} + \sum_{k=1}^{N}\left(1-e^{-q^2\tau_k^2}\right)f_k(q)^2 \quad (5)$$

In the simple case where every atom has the same variance τ^2, we have

$$\mathbb{E}(I_X(q)) = e^{-q^2\tau^2} I_{x^\circ}(q) + \left(1 - e^{-q^2\tau^2}\right) \sum_{k=1}^{N} f_k(q)^2 \tag{6}$$

This result was obtained differently in 1932 by R. W. James [32], as recently rediscovered by P. B. Moore [33], who generalized it to anisotropic motion (*i.e.*, arbitrary diagonal covariance matrix for X). It makes clear that the SAS profile of the thermal ensemble deviates from that of its center structure for momentum transfer values around and above $1/\tau$. For $\tau \ll 1$ Å, this effect is not within the measurable range of q values. However, in systems with large domain movements for which $\tau \gg 1$ Å, this effect becomes of prime importance. The fact that multiple different conformers coexist in solution can then be captured by SAS experiments. Indeed, the SAS curve of x° is then noticeably different from that of the thermal ensemble.

In addition, suppose that our system adopts two different conformations A and B, and that each of these is subject to thermal motions with deviations τ_A and τ_B such that $\tau_A \ll \tau_B$. This can happen, for example, when the system is made of two domains connected by a linker; A would be the state in which the two domains are in contact along a well-defined interaction surface, and B would be when the domains don't interact. Then, assuming no interactions between A and B particles, the average intensity is a weighted sum of the intensities for A and B, each of them given by Eq 6. At low angle, the SAS profile contains information from both conformations. However, because the SAS intensities decay much faster for large τ values, the SAS profile of A will dominate that of B at high angle (assuming the populations of A and B are comparable). Therefore, in SAS, the higher q gets, the more we focus on well-defined conformations. There can be a number of them, but they must be well-defined. On the contrary, continuous conformational variability is more likely only to be noticed at low q values.

Variance and correlation

In any case, because conformations of a thermal ensemble are related, there exist a number of rules that link their SAS profiles together. The SAS profile of one such conformation cannot deviate from Eq 5 in an arbitrary way. This is what we now show, by computing the covariance $\mathbb{V}(I_X(q_i), I_X(q_j)) = \mathbb{E}(I_X(q_i)I_X(q_j)) - \mathbb{E}(I_X(q_i))\mathbb{E}(I_X(q_j))$ between the SAS profile at q_i and q_j. For this purpose, we again use the Debye formula (Eq 1). The expectation of a product of intensities is

$$\mathbb{E}\Big(I_X(q_i)I_X(q_j)\Big) = \sum_{kl}\sum_{mn} f_k(q_i)f_l(q_i)f_m(q_j)f_n(q_j)\mathbb{E}\left(\frac{\sin(q_i d_{kl})}{q_i d_{kl}}\frac{\sin(q_j d_{mn})}{q_j d_{mn}}\right) \tag{7}$$

Then, we notice that

$$\mathbb{E}\left(\frac{\sin(q_i d_{kl})}{q_i d_{kl}}\frac{\sin(q_j d_{mn})}{q_j d_{mn}}\right) = \mathbb{E}\left(\frac{\sin(q_i d_{kl})}{q_i d_{kl}}\right)\mathbb{E}\left(\frac{\sin(q_j d_{mn})}{q_j d_{mn}}\right) \tag{8}$$

when k,l,m,n describe four different atoms. Therefore, the terms that do not cancel out of the covariance calculation are 1) when $k = m$ and $l = n$, *i.e.*, the covariance of a distance with itself, which we call *autocovariance*; and 2) when $k = m$ and $l \neq n$, *i.e.* the covariance between two distances that share a common atom, which we call *cross-covariance*.

First, similar to the calculation of the average intensity, the autocovariance can be given in closed form. It however leads to a formula that is numerically unstable [34]. Second, the cross-covariance cannot be computed in closed form because the probability density function of the

bivariate noncentral χ distribution is not known. Special cases exist for the bivariate noncentral χ^2 probability density function [35] and the characteristic function [36], but the expectation still cannot be calculated.

We therefore seek an approximation to this distribution. A certain number of approaches exist [34, 37], but we use a more direct one (see S1 Text). It is based on a series expansion when all distances are much larger than τ. The bivariate noncentral χ distribution is then approximated as a bivariate normal distribution with mean vector $\mathbf{d}'$ and covariance matrix $\mathbf{\Sigma}$

$$\mathbf{d}' \equiv \begin{pmatrix} d^\circ_{kl} + \dfrac{\tau_k^2 + \tau_l^2}{d^\circ_{kl}} \\ d^\circ_{kn} + \dfrac{\tau_k^2 + \tau_n^2}{d^\circ_{kn}} \end{pmatrix} \quad \mathbf{\Sigma} \equiv \begin{pmatrix} \tau_k^2 + \tau_l^2 & v\tau_k^2 \\ v\tau_k^2 & \tau_k^2 + \tau_n^2 \end{pmatrix} \quad v \equiv \frac{\mathbf{d}^\circ_{kl} \cdot \mathbf{d}^\circ_{kn}}{d^\circ_{kl} d^\circ_{kn}} \tag{9}$$

Using this approximation, and to second order in τ/d°, we can express the correlation and the covariance between the SAS profile at q_i and q_j (see S2 Text)

$$\rho(q_i, q_j) \equiv \frac{\mathbb{V}(I_X(q_i), I_X(q_j))}{\sqrt{\mathbb{V}(I_X(q_i), I_X(q_i))\mathbb{V}(I_X(q_j), I_X(q_j))}} \tag{10}$$

$$\mathbb{V}(I_X(q_i), I_X(q_j)) = V_{\text{auto}}(q_i, q_j) + V_{\text{cross}}(q_i, q_j) \tag{11}$$

$$V_{\text{auto}}(q_i, q_j) \equiv \sum_k f_k(q_i) f_k(q_j) \sum_{l \neq k} f_l(q_i) f_l(q_j) V^\circ_{ij}(d^\circ_{kl}) \tag{12}$$

$$V_{\text{cross}}(q_i, q_j) \equiv \sum_k f_k(q_i) f_k(q_j) \sum_{l \neq k} f_l(q_i) \sum_{n \neq k,l} f_n(q_j) V_{ij}(\mathbf{d}^\circ_{\mathbf{kl}}, \mathbf{d}^\circ_{\mathbf{kn}}) \tag{13}$$

$$V^\circ_{ij}(d^\circ_{kl}) = (\tau_k^2 + \tau_l^2) q_i q_j \sigma(q_i d^\circ_{kl}) \sigma(q_j d^\circ_{kl}) \tag{14}$$

$$V_{ij}(\mathbf{d}^\circ_{\mathbf{kl}}, \mathbf{d}^\circ_{\mathbf{kn}}) = v(\mathbf{d}^\circ_{\mathbf{kl}}, \mathbf{d}^\circ_{\mathbf{kn}}) \tau_k^2 q_i q_j \sigma(q_i d^\circ_{kl}) \sigma(q_j d^\circ_{kn}) \tag{15}$$

$$\sigma(x) \equiv \frac{\mathrm{d}}{\mathrm{d}x} \frac{\sin(x)}{x} = \frac{1}{x} \left(\cos(x) - \frac{\sin(x)}{x} \right) \tag{16}$$

In all cases we studied, the standard deviation (SD) has the characteristic shape of Fig 1 (solid red line, see also S3 Text). The SD starts at zero, consistent with the fact that $I(0)$ is proportional to the number of electrons, and is not impacted by conformational changes. It then quickly reaches a maximum, and then decreases to a plateau. On a relative scale therefore, the standard deviation represents a non-monotonically increasing proportion of the scattered intensity. This finding is consistent with those discussed for the average intensity (Eq 5), in that the conformational diversity is captured at wide angles. We do not expect different hydration models to produce significantly different standard deviations, unless they hydrate different conformers of the ensemble in a different way. However, in the most realistic cases, changes in conformation should cause the solvent shell to rearrange. The water density would therefore be impacted. Consequently, the standard deviation at $I(0)$ could be be nonzero.

We now focus on the the correlation structure of the same SAS profile (Fig 2). In all cases we studied, correlations are strong close to the diagonal, and vanish when points are far apart.

It is also frequent to observe at least one basin with negative correlations. The fact that points that are close together are highly correlated was expected. Indeed, this observation is a simple consequence of the predictable nature of SAS profiles on very short q scales.

Conversely, points that are far apart seem to be largely decorrelated. This fact demonstrates the hierarchical nature of SAS profiles [38]. Being a Fourier transform, the SAS profile describes the shape at low angle. At higher angle, it starts describing the quaternary structure and so forth. What these results suggest, is that SAS compartmentalizes these descriptions. Although individual atoms have a nonzero scattering contribution along the whole range of q values, collectively, a different trend emerges. For example, changes in the quaternary structure that do not modify the overall shape will not affect the onset of the SAS profile.

Another striking feature that can be seen in Fig 2 is that the bandwidth of this correlation matrix varies along the diagonal. Thus, neighboring points will be more or less correlated depending on their absolute position along the SAS profile. That is, the density of independent points along a SAS profile changes as q changes. In information theory, the mutual information of two random variables quantifies how much information one carries on the other. If we take two neighboring points along the SAS profile, their mutual information is

$$I(q_i, q_j) = -\frac{1}{2}\log\left(1 - \rho(q_i, q_j)^2\right) \tag{17}$$

If the mutual information is high, q_i and q_j are strongly related, and consequently the information content of the SAS curve is lower in that region. But $\rho(q_i, q_j)$ is directly related to the bandwidth of the correlation matrix. Therefore, the information content is not uniformly distributed along a SAS profile, and is larger when the bandwidth is smaller.

In all cases studied (see also S3 Text), the bandwidth is large at low q, becomes minimal between $qR_G \sim 3 - 6$ and then broadens again at higher q, suggesting that the information content follows the opposite trends. This result confirms practical observations that the mid-q range ($qR_G \sim 3 - 6$) is the most useful in structure refinement, while high-q, although beneficial, is not as valuable [39].

Extension to correlated motion

The analytical model described until now makes the simplifying assumption that thermalization induces independent random normal displacements for each atom. Such an assumption has strong limitations [33, 40]. In particular, movements in solution are anisotropic, do not follow a normal distribution, and strong correlations between atoms or even protein domains can be expected. To a lesser extent, the bivariate noncentral chi distribution must be approximated to still obtain analytical results. This second-order approximation implies that the resulting covariance formulæare not exact but nonetheless very close, and in any case negligible compared to that of the anisotropic motion. In any case, more realistic representations of thermalization can be obtained with molecular dynamics (MD) simulations. We used the two 1 μs simulations of the lysozyme described by Po-chia Chen and Jochen S. Hub [21], from which we calculated the variance matrix.

Trends in the standard deviations are similar between correlated and independent motion (Fig 3). We see, however, that standard deviations are up to three times larger for correlated motion than for the independent case. They reach 4% of the SAS mean intensity on average, and can go up to 7% at $q = 0.28$ Å^{-1} in this example. These proportions are comparable with the experimental noise level, which commonly ranges from 0.1% to 10% in current experiments. This observation suggests that some structures, which arise naturally through thermal motion, can have a SAS profile that is noticeably different than that of their relatives.

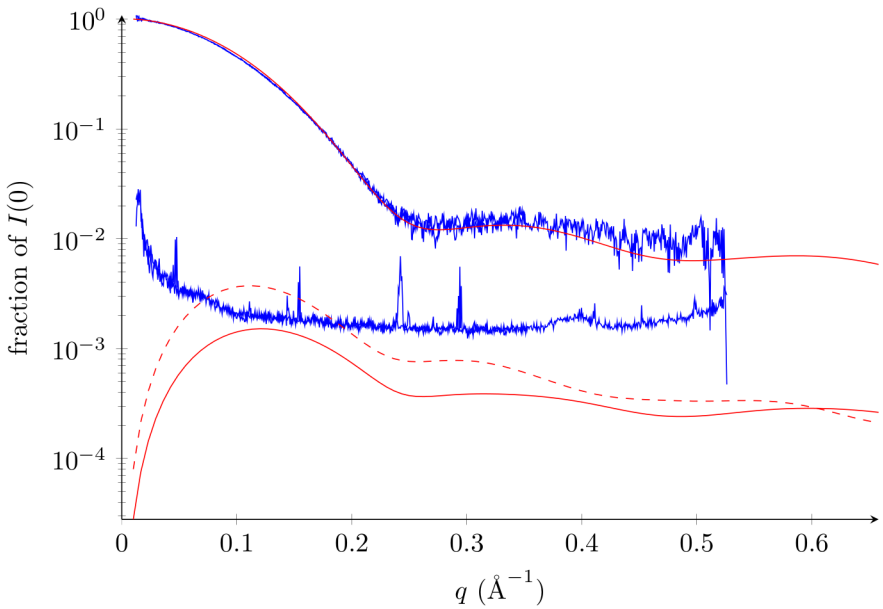


Fig 3. Lysozyme standard deviation compared to signal and noise. Standard deviation of correlated motion (*i.e.*, first MD simulation) is dashed red line (see text for calculation). Standard deviation of independent motion (square root of Eq 11 with $\tau = 0.5$ Å) is bottom solid red line. For reference, the experimental SAXS profile of the lysozyme (top) and its standard error (bottom) are shown in blue (bioisis code LYSOZP). Average SAS profile in the case of independent motion: top solid red.

https://doi.org/10.1371/journal.pone.0177309.g003

They are, however, related through a set of rules which we now describe by looking at the correlations (see also S4 Text). Again, the correlated dataset is very comparable to the independent one. It has approximately the same location for the smallest correlation and the most narrow bandwidth. However, we can observe that 1) the smallest correlation is roughly twice as large, 2) the bandwidth is smaller overall, and 3) new correlation extrema appear between medium and high-q. We do not expect the just described features to change significantly between two hydration models. S4 Text shows the correlation matrix obtained from a second MD simulation. It is reasonable to expect that a change in hydration model would not cause larger differences than those observed between these two simulations.

The depicted correlations can be understood as forming a set of rules that must be satisfied by the SAS profile of any structure within the thermal ensemble. It comes to no surprise, therefore, that the region which has the highest coefficient of variation ($q = 0.28$ Å^{-1}) is the one which is also the most constrained by the correlations. If in some conformers of the thermal ensemble, the SAS profile deviates by 7% from the ensemble SAS profile, then in doing so it must also deviate both at low and high q in a direction that is dictated by the covariances.

As can be seen, the variance grows with the square of the atomic motion (Eqs 14 and 15). For the lysozyme with correlated motion, these variances are comparable to experimental noise levels. For intrinsically disordered proteins in which atomic motion is an order of magnitude larger, this effect dominates the noise, as shown in the case of p15PAF (Fig 4, see also S5 Text) [28]. Therefore, the SAS profile of such a protein is not a static snapshot of one of its conformers, but instead captures its whole conformational complexity.

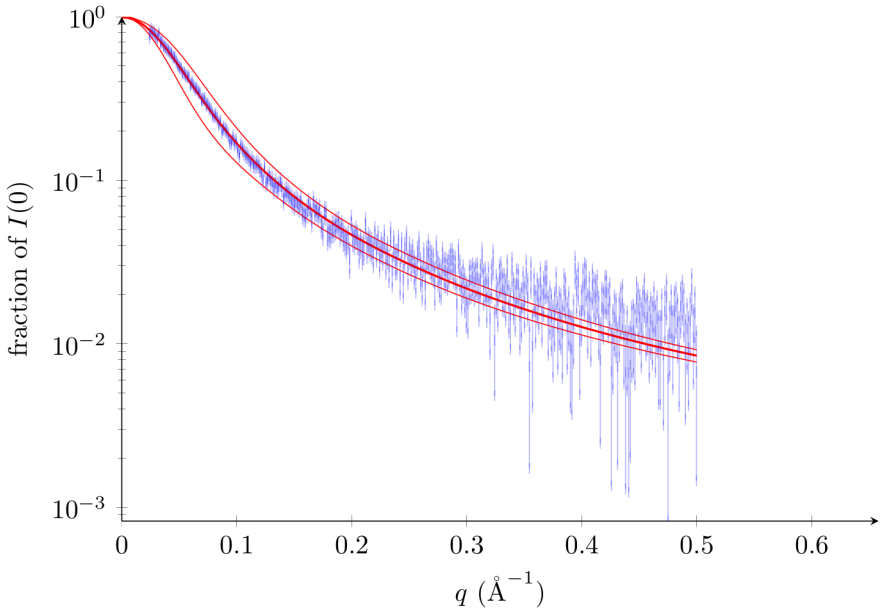


Fig 4. p15PAF experimental SAXS profile and ensemble average SAXS profile. p15PAF profile [28] (PED code PED6AAA) in blue. Ensemble average SAXS profile in thick red (Eq 6). 68% (1σ) confidence interval in red, (Eq $6 \pm \sqrt{\text{Eq } 11}$). The deposited ensemble contains 4939 structures. Individual SAXS profiles were calculated using FoXS with $c1 = 1$ and $c2 = 0.63$.

https://doi.org/10.1371/journal.pone.0177309.g004

Discussion

In this article, we describe the influence of continuous conformational changes on the SAS profile of a protein ensemble. To compute the quantities derived in this article, an atomic or pseudo-atomic model of the protein is needed. They describe how the SAS profile of a structure is modified if it is allowed to be flexible. The resulting SAS profile then contains information on the conformational diversity around that structure. It is however perfectly possible that this ensemble SAS profile be reproduced by a single, different structure. It is up to the modeling expert to determine whether it makes sense to include conformational flexiblity in the modeling or not. However, if the flexible ensemble and the other single structure both fit an experimental profile equally well, Occam's razor would call for a description of the system by the simpler model. Therefore, and as already noted by others, ensemble modelling should only be performed if no satisfactory single conformation can be found.

In the second part of this article, the described SAS covariances are obtained through a long MD simulation. Care must be taken that this simulation is representative of the conformational diversity in solution. Multiple simulations should then yield the same covariance matrix. Unfortunately, even for very long simulations, such as the ones used here, small correlations are very difficult to converge. In our case, the second lysozyme simulation has the same overall covariance structure as described (trends in the variances, behaviour of the bandwidth, location of the global correlation minimum); but there are a number of differences as well: Standard deviations are up to four times larger than the independent case and reach up to 10% of the SAS mean intensity at $q = 0.27$ Å^{-1}. Also, correlations between mid and high-q ranges do not stabilize (see supporting material). We suspect that these differences are mainly due to the fact that the second simulation has an enhanced loop motion [21].

We however hope that it will be possible to measure such a matrix experimentally, alleviating the need for an atomistic model of the protein. Through freezing of the particles in space with cryo-SAXS [41], it should be possible to measure SAS profiles of subsets of the thermal ensemble, and then infer the SAS covariance from them. This approach would work if the solution is sufficiently diluted so that the beam can interact with a small number of molecules, detecting fluctuations from thermodynamic averages. Through freezing in time with the X-ray free electron laser, the coherence of the beam might allow to reconstruct the SAS covariance directly, as already described three decades ago [42–44]. In essence, since for this experiment, the scattering pattern collected on the detector is not radially symmetric, correlations between and within annuli could be related to those of the SAS profile described in this article. We therefore hope that future developments will make the measure of SAS covariances possible.

Conclusion

In this article, we have studied SAS profile correlations. We have shown they reveal the hierarchical nature of SAS profiles. We provided evidence that some portions of the experimental SAS profile are affected by ensemble averaging. Note that the SAS profile correlations described here have nothing in common with those estimated in a recent article, which are correlations of the noise of SAS experiments [45]. We, instead, estimate the correlations that are present within the signal itself.

First, a simple harmonic model of thermal motion allowed to obtain analytical expressions for the correlation between two points in a SAS profile. Second, the analysis of recently published microsecond MD simulations [21] allowed us to see that most trends in the correlations are conserved when thermal motion is modelled with more realism. Third, on the p15PAF structural ensemble [28], SAS profiles of different conformations within that ensemble differ more than the experimental error bar at q. Ensemble averaging can therefore be measured in that region. Last, we believe that these correlations could be measured experimentally with the help of cryo-SAXS or free-electron lasers.

Our developments show that SAS profiles are hierarchical, in the sense that successive regions of the SAS profile are decorrelated. Within these regions however, the knowledge of SAS correlations is essential to correctly describe highly flexible systems, such as intrinsically disordered proteins. We believe that in these systems, the SAS profile alone is not enough to grasp the system's dynamics.

Supporting information

S1 Text. Normal approximation to noncentral χ distributions. Based on the derivations given in [34], this text describes the univariate and bivariate χ distributions, and their approximation when the variance is small.
(PDF)

S2 Text. Calculation of the SAS variance. This text recapitulates and generalizes the derivation given in [34] to the case where each atom has its own variance.
(PDF)

S3 Text. Three additional test cases. SAS mean intensities, standard deviations and correlations are reported for three additional protein test cases.
(PDF)

S4 Text. Lysozyme MD simulations. Lysozyme correlations and standard deviation, as extracted from both MD simulations.
(PDF)

S5 Text. p15PAF. Similar analysis on the ensemble of the p15PAF IDP.
(PDF)

Acknowledgments

YGS would like to acknowledge Dominique Durand, Patrice Vachette and Frédéric Poitevin for fruitful discussion and proofreading of the manuscript, Jochen S. Hub for sharing the lysozyme simulations, and Pau Bernadó for discussion on extension to IDPs. This work was funded by the European Research Commission (Advanced Grant ERC-2011-StG 294809 BayCellS, to MN).

Author Contributions

Conceptualization: YS.

Data curation: YS.

Formal analysis: YS.

Funding acquisition: MN.

Investigation: YS.

Methodology: YS.

Project administration: MN.

Resources: MN.

Software: YS.

Supervision: MN.

Validation: YS.

Visualization: YS.

Writing – original draft: YS.

Writing – review & editing: YS.

References

1. Pérez J, Nishino Y. Advances in X-ray scattering: from solution SAXS to achievements with coherent beams. Curr Op Struct Biol. 2012; 22(5):670–678. https://doi.org/10.1016/j.sbi.2012.07.014 PMID: 22954648
2. Graewert MA, Svergun DI. Impact and progress in small and wide angle X-ray scattering (SAXS and WAXS). Curr Op Struct Biol. 2013; 23(5):748–754. https://doi.org/10.1016/j.sbi.2013.06.007 PMID: 23835228
3. Svergun DI. Small-angle X-ray and neutron scattering as a tool for structural systems biology. Biol Chem. 2010 Jul; 391:737–743. https://doi.org/10.1515/bc.2010.093 PMID: 20482320
4. Guinier A, Fournet G. Small-angle Scattering of X-rays. John Wiley & Sons, Inc.; 1955.
5. Glatter O, Kratky O, editors. Small-angle X-ray Scattering. Academic Press, London; 1982.
6. Beaucage G. Small-Angle Scattering from Polymeric Mass Fractals of Arbitrary Mass-Fractal Dimension. J Appl Cryst. 1996 Apr; 29(2):134–146. https://doi.org/10.1107/S0021889895011605
7. Hammouda B. A new Guinier-Porod model. J Appl Cryst. 2010 05; 43:716–719. https://doi.org/10.1107/S0021889810015773
8. Rambo RP, Tainer JA. Accurate assessment of mass, models and resolution by small-angle scattering. Nature. 2013; 496:477–481. https://doi.org/10.1038/nature12070 PMID: 23619693

9. Rambo RP, Tainer JA. Characterizing Flexible and Intrinsically Unstructured Biological Macromolecules by SAS using the Porod-Debye law. Biopolymers. 2011; 95(8):559–571. https://doi.org/10.1002/bip.21638 PMID: 21509745
10. Moore PB. Small-angle scattering. Information content and error analysis. J Appl Cryst. 1980 Apr; 13(2):168–175. https://doi.org/10.1107/S002188988001179X
11. Taupin D, Luzzati V. Information content and retrieval in solution scattering studies. I. Degrees of freedom and data reduction. J Appl Cryst. 1982 Jun; 15(3):289–300. https://doi.org/10.1107/S0021889882012011
12. Luzzati V, Taupin D. Information content and retrieval in solution scattering studies. II. Evaluation of accuracy and resolution. J Appl Cryst. 1986 Feb; 19(1):39–50. https://doi.org/10.1107/S0021889886090027
13. Koch MHJ, Vachette P, Svergun DI. Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. Q Rev Biophys. 2003 5; 36:147–227. https://doi.org/10.1017/S0033583503003871 PMID: 14686102
14. Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. J Am Chem Soc. 2007; 129(17):5656–5664. https://doi.org/10.1021/ja069124n PMID: 17411046
15. Pelikan M, Hura GL, Hammel M. Structure and flexibility within proteins as identified through small angle X-ray scattering. Gen Physiol Biophys. 2009; 28:174–189. https://doi.org/10.4149/gpb_2009_02_174 PMID: 19592714
16. Yang S, Blachowicz L, Makowski L, Roux B. Multidomain assembled states of Hck tyrosine kinase in solution. Proc Nat Acad Sci USA. 2010; 107(36):15757–15762. https://doi.org/10.1073/pnas.1004569107 PMID: 20798061
17. Rozycki B, Kim YC, Hummer G. SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. Structure. 2011; 19(1):109–116. https://doi.org/10.1016/j.str.2010.10.006 PMID: 21220121
18. Berlin K, Castañeda CA, Schneidman-Duhovny D, Sali A, Nava-Tudela A, Fushman D. Recovering a Representative Conformational Ensemble from Underdetermined Macromolecular Structural Data. J Am Chem Soc. 2013; 135(44):16595–16609. https://doi.org/10.1021/ja4083717 PMID: 24093873
19. Antonov LD, Olsson S, Boomsma W, Hamelryck T. Bayesian inference of protein ensembles from SAXS data. Phys Chem Chem Phys. 2016; 18:5832. https://doi.org/10.1039/C5CP04886A PMID: 26548662
20. Receveur-Bréchot V, Durand D. How Random are Intrinsically Disordered Proteins? A Small Angle Scattering Perspective. Curr Protein Pept Sci. 2012; 13:55–75. https://doi.org/10.2174/138920312799277901 PMID: 22044150
21. Chen P, Hub JS. Validating Solution Ensembles from Molecular Dynamics Simulation by Wide-Angle X-ray Scattering Data. Biophys J. 2014; 107(2):435–447. https://doi.org/10.1016/j.bpj.2014.06.006 PMID: 25028885
22. Theobald DL, Wuttke DS. Accurate Structural Correlations from Maximum Likelihood Superpositions. PLoS Comput Biol. 2008 02; 4(2):e43. https://doi.org/10.1371/journal.pcbi.0040043 PMID: 18282091
23. Schneidman-Duhovny D, Hammel M, Sali A. FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic Acids Res. 2010; 38(suppl 2):W540–W544. https://doi.org/10.1093/nar/gkq461 PMID: 20507903
24. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. Accurate SAXS Profile Computation and its Assessment by Contrast Variation Experiments. Biophys J. 2013; 105(4):962–974. https://doi.org/10.1016/j.bpj.2013.07.020 PMID: 23972848
25. Merzel F, Smith JC. Is the first hydration shell of lysozyme of higher density than bulk water? Proc Nat Acad Sci USA. 2002; 99(8):5378–5383. https://doi.org/10.1073/pnas.082335099 PMID: 11959992
26. Park S, Bardhan JP, Roux B, Makowski L. Simulated x-ray scattering of protein solutions using explicit-solvent models. J Chem Phys. 2009; 130(13):134114–134118. https://doi.org/10.1063/1.3099611 PMID: 19355724
27. Köfinger J, Hummer G. Atomic-resolution structural information from scattering experiments on macromolecules in solution. Phys Rev E. 2013 May; 87:052712. https://doi.org/10.1103/PhysRevE.87.052712 PMID: 23767571
28. De Biasio A, Ibañez de Opakua A, Cordeiro TN, Villate M, Merino N, Sibille N, et al. p15PAF is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins. Biophys J. 2014; 106(4):865–874. https://doi.org/10.1016/j.bpj.2013.12.046 PMID: 24559989
29. Debye P. Zerstreuung von Röntgenstrahlen. Ann Phys. 1915; 351(6):809–823. https://doi.org/10.1002/andp.19153510606

30. Debye P. Interferenz von Röntgenstrahlen und Wärmebewegung. Ann Phys. 1913; 348:49–92. https://doi.org/10.1002/andp.19133480105

31. Waller I. Zur Frage der Einwirkung der Wärmebewegung auf die Interferenz von Röntgenstrahlen. Z Phys. 1923 Dec; 17:398–408. https://doi.org/10.1007/BF01328696

32. James RW. Über den Einfluß der Temperatur auf die Streuung der Röntgenstrahlen durch Gasmoleküle. Phys Z. 1932; 33:737–754.

33. Moore PB. The Effects of Thermal Disorder on the Solution-Scattering Profiles of Macromolecules. Biophys J. 2014; 106(7):1489–1496. https://doi.org/10.1016/j.bpj.2014.02.016 PMID: 24703310

34. Spill YG (2013) Sampling methods development and Bayesian analysis of continuous data. Ph.D. thesis, Université Paris Diderot—Paris 7.

35. Anderson TW, Girshick MA. Some Extensions of the Wishart Distribution. Ann Math Stat. 1944; 15(4): 345–357. https://doi.org/10.1214/aoms/1177731206

36. Tourneret JY, Ferrari A, Letac G. The noncentral wishart distribution: properties and application to speckle imaging. In: IEEE/SP 13th Workshop on Statistical Signal Processing; 2005. p. 924–929.

37. Jensen DR. Gaussian approximation to bivariate rayleigh distributions. J Stat Comput Sim. 1976; 4(4): 259–267. https://doi.org/10.1080/00949657608810129

38. Svergun DI, Koch MHJ. Advances in structure analysis using small-angle scattering in solution. Curr Op Struct Biol. 2002; 12(5):654–660. https://doi.org/10.1016/S0959-440X(02)00363-9 PMID: 12464319

39. Grishaev A, Tugarinov V, Kay LE, Trewhella J, Bax A. Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. J Biomol NMR. 2008 02; 40(2): 95–106. https://doi.org/10.1007/s10858-007-9211-5 PMID: 18008171

40. Makowski L, Gore D, Mandava S, Minh D, Park S, Rodi DJ, et al. X-ray solution scattering studies of the structural diversity intrinsic to protein ensembles. Biopolymers. 2011; 95(8):531–542. https://doi.org/10.1002/bip.21631 PMID: 21462170

41. Meisburger S, Warkentin M, Chen H, Hopkins J, Gillilan R, Pollack L, et al. Breaking the Radiation Damage Limit with Cryo-SAXS. Biophys J. 2013; 104(1):227–236. https://doi.org/10.1016/j.bpj.2012.11.3817 PMID: 23332075

42. Kam Z. Determination of Macromolecular Structure in Solution by Spatial Correlation of Scattering Fluctuations. Macromolecules. 1977; 10(5):927–934. https://doi.org/10.1021/ma60059a009

43. Kirian RA, Schmidt KE, Wang X, Doak RB, Spence JCH. Signal, noise, and resolution in correlated fluctuations from snapshot small-angle x-ray scattering. Phys Rev E. 2011 Jul; 84:011921. https://doi.org/10.1103/PhysRevE.84.059907

44. Mendez D, Lane TJ, Sung J, Sellberg J, Levard C, Watkins H, et al. Observation of correlated X-ray scattering at atomic resolution. Phil Trans R Soc B. 2014; 369(1647):20130315. https://doi.org/10.1098/rstb.2013.0315 PMID: 24914148

45. Franke D, Jeffries CM, Svergun DI. Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra. Nat Methods. 2015; 12:419–422. https://doi.org/10.1038/nmeth.3358 PMID: 25849637