THE PROTEIN SOCIETY

SAXSMoW 2.0: Online calculator of the molecular weight of proteins in dilute solution from experimental SAXS data measured on a relative scale

Vassili Piiadov,[1] Evandro Ares de Araújo,[1] Mario Oliveira Neto,[2] Aldo Felix Craievich,[3] and Igor Polikarpov [1]*

[1]Institute of Physics of São Carlos, University of São Paulo, São Carlos, São Paulo, Brazil
[2]Institute of Biosciences, University Estadual Paulista, Botucatu, São Paulo, Brazil
[3]Institute of Physics, University of São Paulo, São Paulo, São Paulo, Brazil

Received 2 August 2018; Accepted 8 October 2018
DOI: 10.1002/pro.3528
Published online 13 December 2018 proteinscience.org

Abstract: **Knowledge of molecular weight, oligomeric states, and quaternary arrangements of proteins in solution is fundamental for understanding their molecular functions and activities. We describe here a program SAXSMoW 2.0 for robust and quick determination of molecular weight and oligomeric state of proteins in dilute solution, starting from a single experimental small-angle scattering intensity curve, *I*(*q*), measured on a relative scale. The first version of this calculator has been widely used during the last decade and applied to analyze experimental SAXS data of many proteins and protein complexes. SAXSMoW 2.0 exhibits new features which allow for the direct input of experimental intensity curves and also automatic modes for quick determinations of the radius of gyration, volume, and molecular weight. The new program was extensively tested by applying it to many experimental SAXS curves downloaded from the open databases, corresponding to proteins with different shapes and molecular weights ranging from ~10 kDa up to about ~500 kDa and different shapes from globular to elongated. These tests reveal that the use of SAXSMoW 2.0 allows for determinations of molecular weights of proteins in dilute solution with a median discrepancy of about 12% for globular proteins. In case of elongated molecules, discrepancy value can be significantly higher. Our tests show discrepancies of approximately 21% for the proteins with molecular shape aspect ratios up to 18.**

Keywords: SAXS; proteins; molecular weight; on-line calculator

Introduction

Small-angle X-ray scattering (SAXS) is an experimental technique frequently applied to low-resolution structural studies of macromolecules embedded in a homogeneous liquid medium, over a molecular size scale within the 1–100 nm range. The SAXS method allows for investigations of both, well-structured and disordered macromolecules in solution, neither requiring crystallization procedures nor highly elaborate sample preparations. The experimental SAXS intensity associated to a set of proteins of the same

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Conselho Nacional de Desenvolvimento Científico e Tecnológico 140667/2015-6158752/2015-5303988/2016-9405191/2015-4440977/2016-9; Grant sponsor: Fundação de Amparo à Pesquisa do Estado de São Paulo 2014/00769-52015/13684-0.

*Correspondence to: I. Polikarpov, Institute of Physics of São Carlos, University of São Paulo São Carlos São Paulo, Brazil. Email: ipolikarpov@ifsc.usp.br

V. Piiadov and E. Ares de Araújo contributed equally to this work.

nature in dilute solution—after subtracting the parasitic scattering intensity produced by the buffer under the same experimental condition—is proportional to the scattering intensity produced by a single protein averaged over all possible orientations. For a dilute set of proteins with random orientations, the scattering intensity defined in the reciprocal space is isotropic.

In modern laboratory setups and synchrotron radiation-based beamlines, SAXS data are recorded by two-dimensional (2D) detectors. The angular-averaging of 2D detector patterns yields a one-dimensional (1D) scattering intensity as a function of the modulus of the scattering vector, $I(q)$. In order to derive structural information of proteins in solution from SAXS curves, several software packages for data analyses and evaluations, such as ATSAS[1] and SCATTER,[2] are currently available.

In the following, we will focus on SAXS intensity data corresponding to isotropic and dilute sets of monodisperse proteins hydrated by homogeneous buffers. We will also consider that the electron densities of the proteins and the buffers are both constant in space and time. Under this conditions, the relevant parameter associated to SAXS intensity and related to electron densities, known as density contrast, which is defined as $\Delta\rho = \rho_{protein} - \rho_{buffer}$. The value of $\Delta\rho$ only affects the absolute value of the SAXS intensity but not the shape of the scattering intensity curve.

Robust determinations of molecular weight and oligomeric state of proteins in solution are fundamental for understanding their quaternary structure and function. On the other hand, a large number of proteins that are usually studied on SAXS beamlines at most of the existing synchrotron X-ray sources requires quick procedures for achieving quantitative information on molecular weights and oligomeric states of the proteins, during ongoing series of experiments. SAXSMoW 2.0 was designed to achieve these goals.

Results and discussion

User interface and usage

The user interface of SAXSMoW 2.0 is shown in Figure 1. To get started, users must upload the selected *.dat*-file containing the experimental data to be analyzed. A data file is uploaded and processed on the server side automatically and then results are displayed on the same page. "Guinier fitting" section shows the relevant parameters derived from a linear fitting in a Guinier plot, namely the fitting interval in $Å^{-1}$ units, the extrapolated intensity $I(0)$, the radius of gyration R_g and the $q \cdot R_g$ relation associated to the fit.

The second part of the interface regards the choice of the upper limit of integration (q_{max}) for the calculation of the apparent Porod invariant Q'. There are three available options (i) $q_{max} = 8/R_g$, which is the default option, (ii) q_{max} satisfying the condition $\log\frac{I(0)}{I(q)} = 2.25$ and (iii) q_{max} manually selected by the user. For Options 1 and 2, the values of q_{max} are automatically calculated by SAXSMoW 2.0.

The third part of the web interface displays the calculated molecular weight. If the expected molecular weight is known (which can be, for example, computed on the basis of a known amino-acid sequence) and specified, the program also displays the oligomeric state and the discrepancy between calculated molecular weight and the expected value. If necessary, the calculation of molecular weight can be repeated, for example, after manually updating Guinier fitting or varying the upper integration limit.

Figure 1 display the results of calculations shown in the screen using experimental SAXS data from bovine serum albumin (BSA) https://www.sasbdb.org/data/SASDA32/, taken from SASBDB database.[3] Four plots are displayed in Figure 1 showing: (i) experimental SAXS intensity, (ii) Guinier plot $\log I(q)$ versus q^2, (iii) Kratky function $I(q)q^2$ versus q and Porod function $I(q)q^4$ versus q. Finally, a file with all the results can be downloaded.

Precision of molecular weight determinations: globular proteins

We define discrepancies as positive values:

$$D = \left|\frac{m}{m_0} - 1\right| \cdot 100\%, \quad (1)$$

where m and m_0 are calculated and expected molecular weights, respectively. Without the modulus brackets when the distribution is becoming symmetric, we obtain false estimation of mean error of the method because of a mutual compensation of errors from different proteins. For example, positive discrepancy of 50% will erase an impact of negative discrepancy of −50% in a mean estimation of the method precision. Moreover, as we have tested datasets from different proteins with a number of various physical parameters such a shape, weight, measurement conditions, and so forth, we cannot fit the obtained distributions by a Gaussian even without the modulus in Eq. 1. The distribution of discrepancies defined by Eq. 1 is a monotonically decreasing function, thus, we have used the median as a statistical parameter for measuring the narrowness or width of the discrepancy distribution. The distribution for test SAXS datasets measured from globular proteins is displayed in Figures 2 and 3 for both suggested options for q_{max} selection. Notice that the figures do not represents datasets for which automated search of q_{max} is not available.

Figure 2 displays the distribution of calculated molecular weights and their respective discrepancies with respect to their expected values. The molecular weights for most of the evaluated SAXS datasets exhibit discrepancies lower than 10%. This statistics leads us to the conclusion that well-folded and compact proteins, SAXSMoW 2.0 in most cases allow for

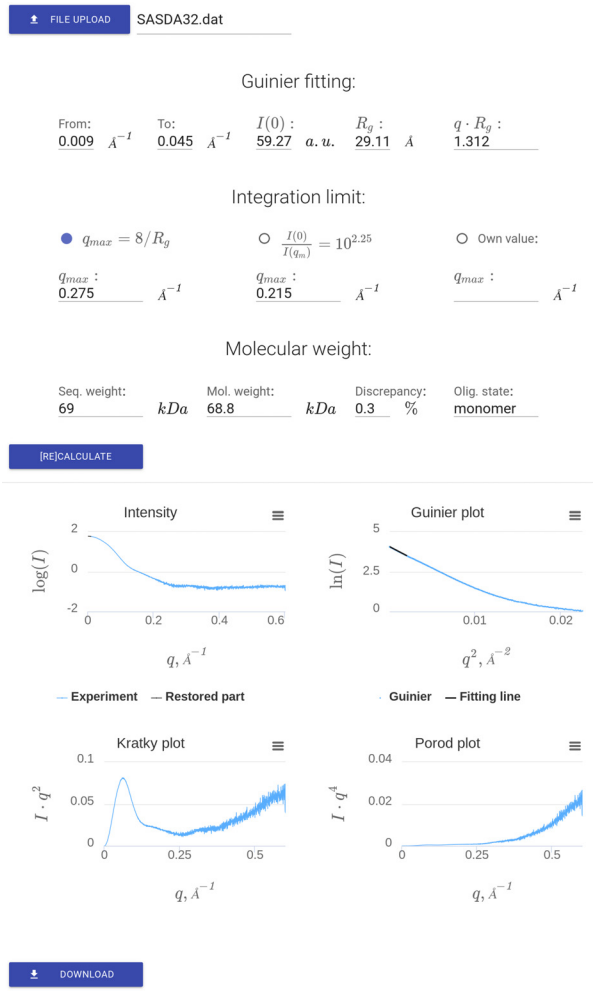

Figure 1. SAXSMoW 2.0 interface displaying results associated to the SASDA32 dataset https://www.sasbdb.org/data/SASDA32/.

determining molecular weights with good accuracy. However, the observed distribution also includes a few outliers with discrepancies larger than 25%.

Figure 3 shows that a number of proteins with a given discrepancy in their molecular weights monotonically decreases for increasing discrepancy values. One can see that the use of the first (default) option, q_{max} = 8/R_g, leads to determinations of the molecular weights with somewhat lower discrepancies, resulting in a more compact discrepancy distribution. This is evidenced by comparing Figure 3(a) with Figure 3(b).

Table I reports statistical features of distributions of discrepancies in molecular weights associated to each suggested options for automatic determination of q_{max} available in SAXSMoW 2.0. Table I reports that for 50% of the dataset, SAXSMoW 2.0 calculates molecular weights with an error smaller than 11.01% when q_{max} is defined as 8/R_g (Option 1) and smaller than 12.25% when q_{max} is defined by the equation $\log\frac{I(0)}{I(q_{max})} = 2.25$ (Option 2). In both cases, the median discrepancy is lower than 12.5%.

Precision of molecular weights determinations: elongated proteins

Eighteen SAXS datasets were selected to establish the influence of non-globularity on the accuracy of the results yielded by SAXSMoW 2.0, which were measured for elongated molecules with estimated aspect ratio ranging from 1.2 to 18. As shown in Figure 4, relative discrepancies between computed molecular weights and those a priori expected, clearly follow an increasing trend as the protein shapes become more elongated. Nevertheless, SAXSMoW 2.0 is quite successful in calculations of the molecular weights of proteins with aspect ratio lower than 8.0 (with discrepancies of about 9.4%) and with aspect ratio of 10 to 18 (with discrepancies of about 21%). Interestingly, q_{max} computed by equation $I(0)/I(q_{max}) = 10^{2.25}$ (Option 2 in Fig. 4) generally achieves better results for molecules having high aspect ratio. Figure 4 exhibits a discrepancy that clearly grows for increasing aspect ratio. In cases for which the degree of elongation of the molecule is a priori known, data plotted in Figure 4 help users to estimate discrepancy boundaries, even if the sequence and weight of the macromolecule is unknown.

The reported analyses indicate that SAXSMoW 2.0 is actually able to determine the molecular weights of a number of proteins with different aspect ratios. However, for SAXS data corresponding to elongated proteins with aspect ratio higher than 10, calculations of the molecular weights become progressively less precise.

Furthermore, in particular cases for which molecular weights are difficult to assess, such as for highly elongated and/or very flexible proteins, Kratky plots do not exhibit a well-defined maximum peak and only show a flat plateau at high q [Supporting Information Fig. S1(a)]. In these more challenging cases, SAXSMoW 2.0 could still be useful for evaluating the oligomeric states of proteins in solution, even though the discrepancy is about 20% [Supporting Information Fig. S1(b,c)].

Comments on particular determinations of molecular weights

As described above, the molecular weights of folded and compact proteins can be determined by using the automatic mode and default option for the upper integration limit, yielding values with average discrepancy of about 10%. Furthermore, in more complex cases, such as those of elongated or flexible proteins, an advanced user can select the q_{max} value manually as it was also implemented in the previous version of SAXSMoW program.[4]

As an example of the program application to scattering data from well-behaved globular proteins, the molecular weights were calculated for different SAXS datasets using available data for a BSA monomer,

BSA dimer, and xylose Isomerase from *Streptomyces rubiginosus*, which are deposited respectively as entries SASDBJ3, SASDBK3,[5] and SASDAB6[6] in SASBDB repository. Using default mode for integration limit determining, obtained molecular weights for monomeric, and dimeric BSA were 64.7*kDa* and 123.2*kDa* with discrepancies of 2% and 7.4%, respectively.

The molecular weight of tetrameric *Streptomyces rubiginosus* was estimated as 157.1*kDa* with a 9.2% discrepancy. This value of the molecular weight agrees with those previously determined by[5] and SASDAB6.[6] However, in cases for which the aspect ratio is approximately 10 or higher, discrepancies in molecular weight estimates are above 10% (see Figure 4) as it is observed for myelin-associated glycoprotein (entry SASDB56)[7] and surface G protein (entry SASDA37)[8] (see Supporting Information Figure S1(b,c)) having aspect ratio of 10 and 18, respectively.

When the upper limit q_{max} is manually selected, SAXSMoW 2.0 consistently leads to a discrepancy in molecular weight up to 10% for globular proteins and larger than 10% for proteins with aspect ratios of 10:1 to 18:1 as shown in Figure 4.

As expected, large discrepancies in molecular weight determinations were observed for unfolded/disordered, metal-depend, and aggregated proteins. This is illustrated, for example, by scattering behavior of human persulfide dioxygenase ETHE1, which is a metal-dependent protein and for its metal-free forms (entries SASDAH7, SASDAJ7, SASDAK7, SASDAL7, SASDAM7, and SASDAN7; https://www.sasbdb.org/project/76/) which are highly elongated. In these cases, linear behaviors of log$I(q)$ versus q^2 at low q are not apparent. Thus, Guinier fitting with acceptable accuracy and extrapolation of $I(q)$ down to q = 0 cannot be done. However, for a metal-bound form of the enzyme (entry SASDAF7), the molecular weight determined by SAXSMoW 2.0 has a discrepancy of 14.7% with respect to the expected molecular weight. Similarly, in the case of the protein ORF 2047.1 from *Pyrococcus furiosus,*[9] which is an unfolded macromolecule; the program has been unsuccessful in determination of its molecular weight.

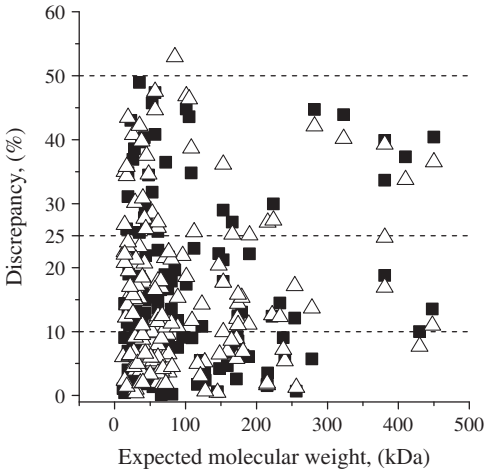

Figure 2. Discrepancy distributions for different expected molecular weights. (■) Option 1: $q_{max} = 8/R_g$; (△) Option 2: derived from equation $I(0)/I(q_{max}) = 10^{2.25}$.

If additional information is a priori known, users can estimate the values of the discrepancy of their results. For example, if the protein is known to be globular and not very small (with molecular weight higher than ~20*kDa*), the relative discrepancy can be considered to be 15% circa. On the other hand, the estimated error also depends on the q_{max} value selected by the user. This, in turn, depends on the statistical quality and other factors. For these reasons our software cannot provide a very accurate estimation of discrepancy. Anyway, we expect that the presented discussion and the results of our several tests referring to proteins with known molecular weight will help users to establish useful estimations of discrepancies for their molecular weights calculated by SASMoW 2.0.

Criteria for selecting the upper limit for integration of Kratky function

After the input of the dataset containing raw SAXS curve, SAXSMoW 2.0 users need to decide which upper integration limit, q_{max}, to select for the calculation of the truncated integral of the Kratky function (Eq. 6).

Our suggestion is to select the default option ($q_{max} = 8/R_g$), which is widely used in several packages for analyses of SAXS results. It is noteworthy that the SAXS intensity $I(q)$ up to this upper q-limit contains most of the relevant structural information associated to strictly homogeneous particles. This implies that the comparatively weak effects from molecular flexibility and density fluctuations are expected to strongly affect the Kratky function mainly above ~$8/R_g$. A strong contribution to the integral of the $I(q)q^2$ function above $q = 8/R_g$ can clearly be seen in the example of Kratky plot corresponding to the SASDA32 data set shown in Figure 1. The same behavior is apparent in Porod plots of SAXS curves of many other proteins.

In some cases, the first option for q_{max} may lie outside the available q_{max} = 0 range (0.1 Å^{-1} <q_{max} < 0.5 Å^{-1}) over which the A and B parameters of the linear function Eq. 9 are defined. For these SAXS curves, the second option for q_{max} can be tested. If selecting the second option yields q_{max} > 0.5 Å^{-1}, the suggested choice is to use q_{max} = 0.5 Å^{-1}.

For many proteins with the molecular weights below 20 *kDa*, the upper limit q_{max} for the integration of the Kratky function is higher than 0.5 Å^{-1}. For the analysis of these small proteins, the use of the maximum value q_{max} = 0.5 Å^{-1} is advisable.

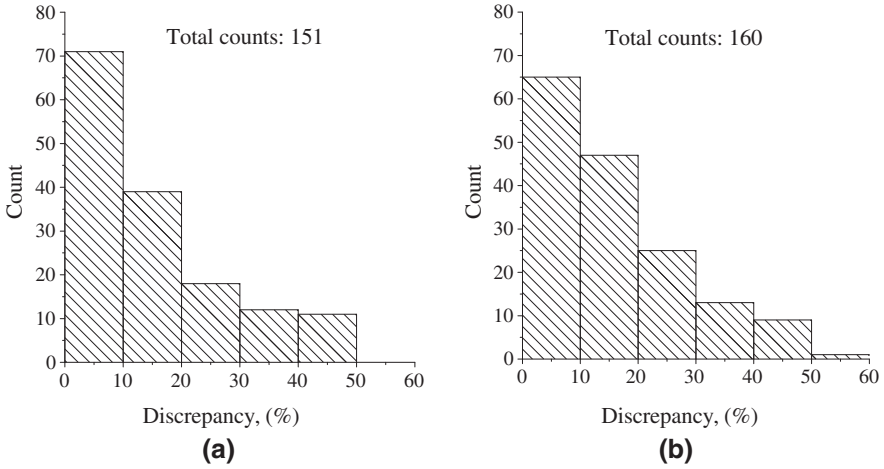

Figure 3. Discrepancy distributions in molecular weights computed for a set of globular proteins using different q_{max} values: (a) $q_{max} = 8/R_g$ and (b) q_{max} from equation $\log\frac{I(0)}{I(q_{max})} = 2.25$.

Alternatively, users may opt for a manual mode of q_{max} selection. SAXSMoW 2.0 allows for choosing any q_{max} value between 0.1 $Å^{-1}$ and 0.5 $Å^{-1}$. In this case, users should avoid to select too low q_{max} to avoid strong truncation of the $I(q)q^2$ function and keep q_{max} below the high q-range over which the contribution to the integral Q' from density fluctuations is high.

Final remarks

SAXSMoW 2.0 is a user-friendly program for robust and quick online determinations of the molecular weight of proteins in dilute solution from experimental SAXS intensity data collected on a relative scale. This program builds up on its previous version,[4] which was widely applied during the last decade. The SAXSMoW 2.0 exhibits new features with respect to the previous version, namely:

- Input of background-subtracted SAXS intensity curves without the need to use auxiliary packages;
- display of experimental data as $I(q)$, $I(q)q^2$ and $I(q)q^4$ for visual examination;
- automatic Guinier fitting of the experimental SAXS intensity at low q, calculation of the molecular radius of gyration, R_g, and determination of $I(0)$ by extrapolation of SAXS curve down to $q = 0$;
- suggestions of two options of upper integration limits for calculation of the truncated integrals of Kratky functions which allow for quick and automatic determinations of molecular weights;
- possibility for calculations of molecular weight by selecting any value for the upper integration limit q_{max} within the $0.1Å^{-1} < q_{max} < 0.5Å^{-1}$ range to compute the apparent molecular volume.

Our test analyses of many openly available SAXS datasets indicate that SAXSMoW 2.0 allows for determining molecular weights with a median discrepancy lower than 12% for globular (i.e., not very elongated) and homogeneous proteins. For elongated proteins having aspect ratios up to 18 and highly flexible proteins, the discrepancies are much higher.

In this article, we have discussed several practical uses of SAXSMoW 2.0 for determinations of molecular weights of a number of proteins in dilute solutions. All these analyzed data are related to different results from SAXS measurements. Needless to say, this program can also be applied to the experimental data of small-angle scattering of neutrons (SANS).

The program SAXSMoW 2.0 is now fully implemented online and freely available at http://saxs.ifsc.usp.br webpage.

Procedures for determination of protein molecular weight

The molecular weight of proteins in solution can be determined from experimental SAXS function from the value of the intensity in absolute scale extrapolated to zero angle, $I(0)$. Another method uses the extrapolated intensity $I(0)$ in relative scale and further comparison with $I(0)$ from a standard protein with known molecular weight.[10] However, these methods exhibit several sources of errors that often introduce systematic bias in the assessment of a protein molecular weight.[11]

Alternative method for determination of the molecular weight of protein in a dilute solution, that

Table I. *Statistics on distributions of discrepancy* D *for globular proteins set.*

Option 1: $q_m = 8/R_g$		
Minimum	Median	Maximum
0.08	11.01	48.95
Option 2: q_{max} by eq. $I(0)/I(q_{max}) = 10^{2.25}$		
Minimum	Median	Maximum
0.33	12.25	52.90

applies to SAXS data collected on a relative scale, that has recently been incorporated in the DatPorod program from ATSAS package.[1] This method is based on the determination of the quotient between the extrapolated SAXS intensity, $I(0)$, and the Porod invariant, that is, the integral of the Kratky function,

$$Q = \int_0^\infty I(q)q^2 dq. \quad (2)$$

The $I(0)$ value in relative units is determined by extrapolation down to $q = 0$ from the linear low-q range of Guinier plots ($\log I(q)$ vs q^2) while the determination of the integral Q is more difficult because it requires the extrapolation of $I(q)$ up to $q = \infty$ that is performed up to the high q-range, where the scattering intensity is usually low and the relative statistical error is high. DatPorod automatically subtracts the contribution to SAXS intensity originating from the effects of protein flexibilities and from the fluctuations in density due to minor heterogeneities in the protein structures. This procedure follows by a further extrapolation of SAXS intensity up to $q = \infty$ by applying Porod equation $I(q) \propto q^{-4}$. After computing $I(0)$ and Q, the molecular volume of the protein is determined as

$$V = 2\pi^2 \frac{I(0)}{Q}. \quad (3)$$

Finally, the product of the calculated protein volumes and the known mass density yields their molecular weight. DATPorod determines molecular weights of proteins with a relative discrepancy of circa 20%.[12]

Furthermore, reference [13] describes another method for determination of molecular weights based on the calculation of the volume of correlation V_c and a molecular mass estimator Q_R, both derived from experimental SAXS intensity curves. Differently than DatPorod program,[1] this method does not calculate the Porod invariant Q. We did not carry out a critical comparison of results derived by using the approaches described in references,[1,13] and that of SAXSMoW 2.0.

The previous version of SAXSMoW[4] was developed for estimation of the molecular weights of proteins in dilute solution starting from a single SAXS curve measured at a relative scale. Instead of calculating the "true invariant", Q, as DatPorod does, this program determines a truncated or "apparent" Porod invariant, Q'. SAXSMoW is a simple, fast, and relatively precise method for determinations of the molecular weight of proteins in dilute solutions.[14] The program has been widely used during the last decade to determine the molecular weights of many proteins with different shapes and forms, including globular, elongated, flexible, and glycosylated proteins, and also protein complexes.[8,15–23]

SAXSMoW is also currently used as an accurate tool for quick diagnosing of protein molecular weights, such as reported for UltraScan-SOMO SAXS pipeline[24] at SOLEIL synchrotron SWING beamline, for BioXTAS RAW SAXS pipeline[25] at the Cornell High Energy Synchrotron Source bioSAXS CHESS beamline, for BL16B1 SAXS beamline at the Shanghai Synchrotron Radiation Facility[26] and summarized in the workflow for determinations of molecular weights and quaternary structure of proteins in solution.[27]

Here we describe the SAXSMoW 2.0 program, a new version of SAXSMoW available at http://saxs.ifsc.usp.br/, which is a web-based utility for processing SAXS data. This web application can be used online without the need of downloading. The input is a one-dimensional SAXS intensity text file (".dat" file) containing at least two columns: The modulus of the scattering vector q in $Å^{-1}$ or nm^{-1}, which automatically will be converted in $Å^{-1}$, and the scattering intensity, $I(q)$, in arbitrary or relative units. All other columns in the uploaded file, if any, will be discarded.

The program automatically performs Guinier fitting, computes the molecular radius of gyration R_g, generates Kratky $I(q)q^2$ and Porod $I(q)q^4$ plots, suggests q-intervals for the calculation of the Q'-invariant, and determines the molecular weight of protein from the SAXS data recorded in experiments.

The input of the previous version of the SAXSMoW program[4] is the regularized intensity function obtained by indirect Fourier transform (IFT) of the scattering intensity. This transformation was performed using the GNOM program of ATSAS package.[12] The newly developed SAXSMoW 2.0 program directly applies the Guinier extrapolation to raw experimental SAXS data (after subtracting the scattering intensity from the buffer), without a need to use ATSAS package.

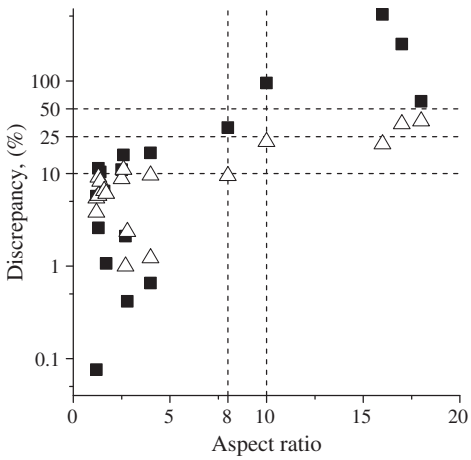

Figure 4. Discrepancies in molecular weights associated to elongated proteins with different aspect ratios for both suggested options for q_{max} values. Option 1 (■) $q_{max} = 8/R_g$. Option 2 (△) derived from equation $I(0)/I(q_{max}) = 10^{2.25}$.

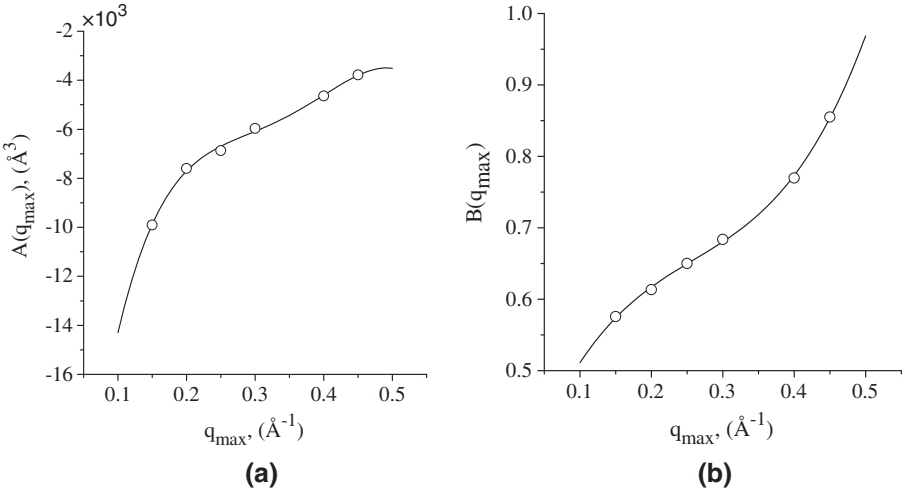

Figure 5. Plots of polynomials that define (a) $A(q_{max})$ and (b) $B(q_{max})$, for all q_{max} values from 0.1 Å^{-1} to 0.5 Å^{-1}.

The method applied by SAXSMoW 2.0

Computation of the molecular radius of gyration and extrapolation of SAXS intensity to q = 0

The radius of gyration of homogeneous particles with a constant electron density defined as $R_g = [(1/V)\int r^2 dv]^{1/2}$ characterizes their size and compactness. This parameter can be determined from SAXS data in several different ways. SAXSMoW 2.0 utilizes a method based on Guinier approximation for the scattering intensity, which applies to SAXS curves at low q.[28] Guinier approximation can be written as

$$\ln I(q) = \ln I(0) - q^2 R_g^2/3. \quad (4)$$

Thus, R_g is determined from the slope of a straight line that asymptotically (at low q) fits the experimental Guinier ($\log I(q)$ vs q^2) plot.

The accuracy of R_g determined from Guinier analysis depends on a number of factors. First of all, interference effects on SAXS curves due to spatial correlation of protein positions may strongly affect the low-q region of the scattering curve. Protein aggregation is one of the sources of systematic errors in data analysis, which leads to a significant increase in the scattering intensity at very small angles. The scattering intensity from aggregates overlaps with the signal from the remaining non-aggregated molecules and changes the shape of the scattering curve. The scattering intensity from aggregates overlaps the signal from the remaining non-aggregated molecules and changes the shape of the scattering curve. Depending on the molecular electrical charges and pH of the solution, concentrated solutions may exhibit effects of intermolecular repulsion. Similarly to the effects of molecular aggregation, those of interparticle repulsion may also significantly affect the shape of the SAXS curves. Thus, the determination of molecular weights from SAXS curves associated to concentrated solutions is more difficult and less precise than using SAXS curves for dilute solutions. In order to eliminate correlations or interference effects in the SAXS curves, the proteins in solution should be studied under dilute conditions. To accomplish this, SAXS measurements are carried out for several protein concentrations and the results are extrapolated to zero concentration. The experience gained by frequent users of SAXS beamlines associated to synchrotron X-ray sources, usually allows them a priori estimations of the required protein concentrations for achieving dilute conditions, without the time-consuming extrapolation procedure described above. Protein concentrations in typical dilute solutions are of the order of a few mg/mL.

SAXSMoW 2.0, by default, performs Guinier fitting automatically, but the program also offers an option for manually defining values for q-range of fitting interval. The implemented strategy of Guinier fit search is based on testing of all possible Guinier fits and selecting the one with the best fit, within the range $R_g \cdot q < 1.3$ by combined criteria of Pearson correlation coefficient and a length of the fitting q^2-range. The area of search is limited by q value of 0.15 Å^{-1}. Depending on the resolution of the experimental curve, a minimal q^2-range length of acceptable fit was defined as $3 \cdot 10^{-3}$ Å^{-1} or q-range corresponding to the first five experimental points of the dataset if it corresponds to a higher resolution. Moreover, imported data having more than 10^6 points is limited to this number of points.

For globular particles, analysis of the accuracy of R_g calculations obtained from experimental SAXS data in the interval $q \cdot R_g < 1.3$ results in a systematic error lower than a few percent. Thus, small errors in this interval allow for a reliable determination of R_g. Over the $q \cdot R_g < 1.5$ interval, the deviation can reach 20–30%, while over the $q \cdot R_g > 2$ region, this approximation becomes highly imprecise.[28,29] SAXSMoW 2.0 checks $q \cdot R_g$ values for each tested fitting line and discards the ones for which the mentioned

relation is not satisfied. A final set of possible fits is first selected by maximizing the Pearson coefficient and, second, by maximizing the length of the fit range, in such a way that considering two fits with similar Pearson coefficients, the fit with a larger linearity range is selected.

Theoretically, the best Guinier approximation should be the one that exhibits the highest Pearson coefficient. However, this allows for cases situations in which, for example, the fitting line with Pearson coefficient 0.995 and fit interval length of five experimental points would be preferred to another fitting straight line with a Pearson coefficient equal to 0.994 and a total length of 100 points. Obviously, choosing the first option would be a wrong decision because Pearson coefficients 0.995 and 0.994 indicate similar (equal) quality of the approximation in the statistical sense. On the other hand, fittings over higher numbers of experimental points provide a more robust result, which is less sensitive to eventual local artifacts of data set. Thus, in such cases, the second option is selected. To implement this, fitting lines with differences in Pearson coefficients smaller than 0.001 are assumed to be with same fitting quality and thus, in these cases, the line with a larger fitting interval is selected. After application of such decision filter, the SAXSMoW 2.0 algorithm selects a Guinier fitting with a higher Pearson coefficient and assumes it as the best fitting line to be used for $I(0)$ calculation.

Thus, regardless of the protein internal structure, the SAXSMoW 2.0 analysis of scattering curves at low q yields R_g and $I(0)$, which depend on the size and compactness of the particle, and on the amount of scattering matter, respectively.

Determinations of the protein volume and molecular weight

For the determination of the molecular volume and molecular weight, SAXSMoW 2.0 starts from the calculation of the apparent protein volume, V', derived from the following equation:

$$V' = 2\pi^2 \frac{I(0)}{Q'}, \tag{5}$$

where $I(0)$ is the SAXS intensity extrapolated to $q = 0$ which is derived from the linear fitting procedure described in the previous sub-section and Q' is named apparent Porod invariant, which is the truncated integral of the Kratky $I(q)q^2$ function from $q = 0$ up to a selected q_{max}:

$$Q' = \int_0^{q_{max}} I(q)q^2 dq. \tag{6}$$

Notice that V' and Q' are named as apparent volume and apparent Porod invariant, respectively, because the determination of their true values require integration of the Kratky function from $q = 0$ up to $q = \infty$.

The first choice for upper limit of integration used by SAXSMoW 2.0 is given by

$$q_{max} \sim 8/R_g, \tag{7}$$

which corresponds to the estimated maximum value of q which contains relevant information associated with perfectly homogeneous particles. This q_{max} value is often used as, for example, in ATSAS software.[12]

Another option for determining of q_{max} is suggested in[29,30]:

$$\log\frac{I(0)}{I(q)} \sim 2..2.5. \tag{8}$$

Thus, the equation $\log[I(0)/I(q)] = 2.25$ was implemented as a second option of q_{max} in SAXSMoW 2.0. Value of 2.25 was chosen as an average value for the interval from Eq. 8.

The next step of SAXSMoW 2.0 is to establish the relations between the true protein volume, V, and the apparent protein volumes associated to different q_{max} values, $V'(q_{max})$. For this purpose, CRYSOL program[1] is used for determinations of the SAXS functions of a large number (1148) of proteins with known 3D high-resolution structures deposited in the PDB. The integrals of Kratky functions truncated at different q_{max} values and the values of $I(0)$ are determined for all selected proteins, which allow for the calculation of their apparent volumes $V'(q_{max})$. Moreover, the true volumes, V, of all selected proteins are easily computed from their known high-resolution structures. Thus, as described with more detail in a previous work,[4] the true protein volumes V was found to exhibits dependences on the apparent volume V' for all selected q_{max}, given by

$$V = A + B \cdot V'. \tag{9}$$

The A and B coefficients were determined in the first version of SAXSMoW for several q_{max} values corresponding to the $V(V')$ function built up by starting from the known high-resolution structures of 1148 selected proteins downloaded from the PDB.[4]

The linear Eq 9 is used for determining true volumes of proteins from their apparent volume computed from experimental SAXS curves truncated at one of the q_{max} values for which the coefficient A and B were reported in[4]. In SAXSMoW 2.0, coefficients A and B are interpolated over the whole q-range, from $q = 0.1$ Å^{-1} up to 0.5 Å^{-1}, by the following polynomials:

$$\begin{aligned} A[\text{Å}^3] &= -2.114\cdot10^6 q_{max}^4 + 2.920\cdot10^6 q_{max}^3 - \\ &\quad -1.472\cdot10^6 q_{max}^2 + 3.349\cdot10^5 q_{max} - 3.577\cdot10^4 \\ B &= 12.09 q_{max}^3 - 9.39 q_{max}^2 + 3.03 q_{max} + 0.29, \end{aligned} \tag{10}$$

in which $[q_{max}] = Å^{-1}$. A and B values for different q and their polynomial approximation given by 9 are shown in Figure 5.

Finally, the molecular weight of proteins is calculated from their volume V by

$$MW[kDa] = \frac{\rho_m[g/cm^3]V[cm^3]}{1.662 \cdot 10^{-21}[g/kDa]} \quad (11)$$

where $\rho_m = 1.37 g/cm^3$ is the mass density of proteins. This value is assumed to be the same for all globular proteins.[31,32]

Testing SAXSMoW 2.0 on experimental datasets

In order to evaluate the accuracy in the results yielded by SAXSMoW 2.0, we have applied it to a number of SAXS curves downloaded from SASBDB[3] and BIOISIS http://www.bioisis.net databases. These databases contain publicly available experimental SAXS data of many proteins, and protein complexes such as protein–protein, protein-DNA, and protein-RNA.

We have excluded in the selection of SAXS datasets those corresponding to (i) aggregated or not purified proteins, (ii) protein-DNA/RNA complexes, (iii) partially folded/unfolded or very disordered proteins, and (iv) metalloproteins. Furthermore, 175 datasets containing SAXS intensity curves corresponding to well-folded proteins were selected for analysis (Supporting Information Table S1). Notice that the expected molecular weights of all selected proteins are known.

Furthermore, in order to better understand the influence of non-globularity on the quality of the SAXSMoW 2.0 calculations, we have selected 18 experimental SAXS curves from proteins having aspect ratios ranging between 1.2 and 18. The aspect ratios were computed by dividing the lengths of two main axes in a spheroid approximation for the protein shape, based on its 3D model (Supporting Information Table S2). Both datasets were processed by SAXSMoW 2.0 in automatic mode using both suggested options for q_{max}, as described in the previous subsection.

Acknowledgments

This research was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) via grant 2015/13684-0 and 2014/00769-5; and by Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq) via grants 405191/2015-4, 140667/2015-6, 158752/2015-5, 303988/2016-9 and 440977/2016-9.

Conflict of Interest

All the authors declare no conflict of interest.

References

1. Franke D, Petoukhov M, Konarev P, Panjkovich A, Tuukkanen A, Mertens H, Kikhney A, Hajizadeh N, Franklin J, Jeffries C, et al. (2017) Atsas 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. J Appl Crystallogr 50(4): 1212–1225.
2. R. Rambo, Scatter, a java-based program for biosaxs. http://bioisis.net/, 2015.
3. Valentini E, Kikhney AG, Previtali G, Jeffries CM, Svergun DI (2014) Sasbdb, a repository for biological small-angle scattering data. Nucleic Acids Res 43: 357–363.
4. Fisher H, de Oliveira Neto M, Napolitano H, Polikarpov I, Craievich A (2010) Determination of the molecular weight of proteins in solution from a single small-angle x-ray scattering measurement on a relative scale. J Appl Crystallogr 43:101–109.
5. Jeffries CM, Graewert MA, Blanchet CE, Langley DB, Whitten AE, Svergun DI (2016) Preparing monodisperse macromolecular samples for successful biological small-angle x-ray and neutron-scattering experiments. Nat Protoc 11(11):2122–2153.
6. Franke D, Jeffries CM, Svergun DI (2015) Correlation map, a goodness-of-fit test for one-dimensional x-ray scattering spectra. Nat Methods 12(5):419.
7. Pronker MF, Lemstra S, Snijder J, Heck AJ, Thies-Weesie DM, Pasterkamp RJ, Janssen BJ (2016) Structural basis of myelin-associated glycoprotein adhesion and signalling. Nat Commun 7:13584.
8. Gruszka DT, Whelan F, Farrance OE, Fung HK, Paci E, Jeffries CM, Svergun DI, Baldock C, Baumann CG, Brockwell DJ, et al. (2015) Cooperative folding of intrinsically disordered domains drives assembly of a strong elongated protein. Nat Commun 6:7271.
9. Hura GL, Menon AL, Hammel M, Rambo RP, Poole Ii FL, Tsutakawa SE, Jenney FE Jr, Classen S, Frankel KA, Hopkins RC, et al. (2009) Robust, high-throughput solution structural analyses by small angle x-ray scattering (SAXS). Nat Methods 6(8):606.
10. Orthaber D, Bergmann A, Glatter O (2000) SAXS experiments on absolute scale with kratky systems using water as a secondary standard. J Appl Crystallogr 33(2): 218–225.
11. Trewhella J, Duff AP, Durand D, Gabel F, Guss JM, Hendrickson WA, Hura GL, Jacques DA, Kirby NM, Kwan AH, et al. (2017) 2017 publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution: an update. Acta Crystallographica Sec D Struct Biol 73(9):710–728.
12. Petoukhov MV, Franke D, Shkumatov AV, Tria G, Kikhney AG, Gajda M, Gorba C, Mertens HD, Konarev PV, Svergun DI (2012) New developments in the atsas program package for small-angle scattering data analysis. J Appl Crystallogr 45(2):342–350.
13. Rambo RP, Tainer JA (2013) Accurate assessment of mass, models and resolution by small-angle scattering. Nature 496(7446):477–481.
14. Guttman M, Weinkam P, Sali A, Lee KK (2013) All-atom ensemble modeling to analyze small-angle x-ray scattering of glycosylated proteins. Structure 21(3):321–331.
15. Glauninger H, Zhang Y, Higgins KA, Jacobs AD, Martin JE, Fu Y, Coyne HJ 3rd, Bruce KE, Maroney MJ, Clemmer DE, et al. (2018) Metal-dependent allosteric activation and inhibition on the same molecular scaffold: the copper sensor copy from streptococcus pneumoniae. Chem Sci 9(1):105–118.
16. Chang A, Abderemane-Ali F, Hura GL, Rossen ND, Gate RE, Minor DL Jr (2018) A calmodulin c-lobe ca2+

–dependent switch governs kv7 channel function. Neuron 97(4):836–852.
17. Kadowaki MA, Higasi P, Godoy MO, Prade RA, Polikarpov I (2018) Biochemical and structural insights into a thermostable cellobiohydrolase from myceliophthora thermophila. FEBS J 285(3):559–579.
18. Ferreira FM, Oliveira LC, Germino GG, Onuchic JN, Onuchic LF (2011) Macromolecular assembly of polycystin-2 intracytosolic c-terminal domain. Proc Natl Acad Sci 108(24):9833–9838.
19. Zheng J, Gay DC, Demeler B, White MA, Keatinge-Clay AT (2012) Divergence of multimodular polyketide synthases revealed by a didomain structure. Nat Chem Biol 8(7):615–621.
20. Carter L, Kim SJ, Schneidman-Duhovny D, Stöhr J, Poncet-Montange G, Weiss TM, Tsuruta H, Prusiner SB, Sali A (2015) Prion protein–antibody complexes characterized by chromatography-coupled small-angle x-ray scattering. Biophys J 109(4):793–805.
21. Hunkeler M, Stuttfeld E, Hagmann A, Imseng S, Maier T (2016) The dynamic organization of fungal acetyl-coa carboxylase. Nat Commun 7:11196.
22. de Araújo EA, Manzine LR, Piiadov V, Kadowaki MAS, Polikarpov I (2017) Biochemical characterization, low-resolution saxs structure and an enzymatic cleavage pattern of blcel48 from bacillus licheniformis. Int J Biol Macromol 111:302–310.
23. Liberato MV, Silveira RL, Prates ÉT, De Araujo EA, Pellegrini VO, Camilo CM, Kadowaki MA, Neto MO, Popov A, Skaf MS, et al. (2016) Molecular characterization of a family 5 glycoside hydrolase suggests an induced-fit enzymatic mechanism. Sci Rep 6:23473.
24. Brookes E, Vachette P, Rocco M, Pérez J (2016) US-SOMO HPLC-SAXS module: dealing with capillary fouling and extraction of pure component patterns from poorly resolved SEC-SAXS data. J Appl Crystallogr 49(5):1827–1841.
25. Hopkins JB, Gillilan RE, Skou S (2017) Bioxtas raw: improvements to a free open-source program for small-angle x-ray scattering data reduction and analysis. J Appl Crystallogr 50(5):1545–1553.
26. Zeng J, Bian F, Wang J, Li X, Wang Y, Tian F, Zhou P (2017) Performance on absolute scattering intensity calibration and protein molecular weight determination at bl16b1, a dedicated saxs beamline at ssrf. J Synchrotron Radiat 24(2):509–520.
27. Korasick DA, Tanner JJ (2018) Determination of protein oligomeric structure from small-angle x-ray scattering. Protein Sci 27:814–824.
28. Guinier A (1939) La diffraction des rayons x aux tres petits angles: applications a l'etude de phenomenes ultramicroscopiques. Ann Phys 11:161–237.
29. Feigin L, Svergun D. *Structure analysis by small-angle X-ray and neutron scattering*. LLC: Springer Science +Business Media, 1987.
30. Kayushina R, Rolbin Y, Feigin L (1974) Determination of the volume of biological macromolecule by mean of small-angle x-ray scattering. Sov Phys Crystallogr 19: 1161–1165.
31. Gekko K, Noguchi H (1979) Compressibility of globular proteins in water at 25°C. J Phys Chem 83(21): 2706–2714.
32. Squire PG, Himmel ME (1979) Hydrodynamics and protein hydration. Arch Biochem Biophys 196:165–177.