JOURNAL OF APPLIED CRYSTALLOGRAPHY

ISSN: 1600-5767
journals.iucr.org/j

Calculation of pair distribution functions from small-angle X-ray scattering protein data by direct transform

John Badger

J. Appl. Cryst. (2025). **58**, 119–127

JOURNAL OF APPLIED CRYSTALLOGRAPHY
ISSN 1600-5767

Calculation of pair distribution functions from small-angle X-ray scattering protein data by direct transform

John Badger*

DeltaG Technologies, 4168 Stephens St., San Diego, CA 92103, USA. *Correspondence e-mail: dgtscientific@gmail.com

Received 16 August 2024
Accepted 2 December 2024

Edited by J. Hajdu, Uppsala University, Sweden and The European Extreme Light Infrastucture, Czechia

Keywords: small-angle X-ray scattering; SAXS data analysis; pair distribution function; protein structure analyses.

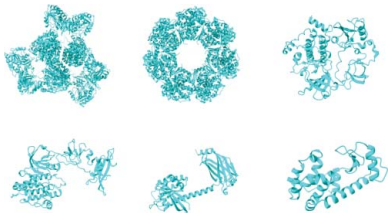
In a small-angle X-ray scattering analysis of protein molecules in solution the calculation of the pair distribution function, $P(r)$, is invariably performed by an indirect Fourier transform. This approach models a $P(r)$ to fit the available intensity data, $I(q)$. The determination of $P(r)$ via a direct transform from $I(q)$ has been dismissed as unworkable since the range of q that is experimentally measured is necessarily incomplete. Here, it is shown that, provided suitable measures are taken to estimate unmeasured low-resolution data and avoid a sharp data truncation at the high-resolution data limit, the appearance of significant artifacts in the resulting $P(r)$ may be circumvented. Using several examples taken from the Small Angle Scattering Biological Data Bank, it is demonstrated that the $P(r)$ obtained by a direct transform are in close agreement with the $P(r)$ obtained using the popular indirect transform program *GNOM*.

1. Introduction

As a result of technological advances in X-ray sources, beam optics, instrumentation and detector hardware, it is now routinely possible to measure accurate X-ray scattering data, $I(q)$, from protein molecules in solution. Although small-angle X-ray scattering (SAXS) data sets are often limited to a resolution of $q \sim 0.3$ Å^{-1}, the shapes of the intensity curves are sensitive to small structural differences between related samples, for example due to variations in environmental conditions or the presence of different cofactors. Data collection from several dozen samples within a single shift at a synchrotron X-ray source is readily achievable.

Since SAXS data are measured in reciprocal space and the dynamic range of measurements across the resolution range is very large, it is generally difficult to infer structural information from a visual inspection of $I(q)$. The analysis of pair distribution functions, $P(r)$, obtained from a transform of the X-ray scattering data, $I(q)$, is a widely used approach for investigating molecular structure in many branches of solid- and liquid-state physics (Billinge, 2019). In the case of protein molecules in solution, the pair distribution function is essentially a histogram of interatomic distances within the protein molecule that encodes information on the size and shape of the structure. The calculation of the pair distribution function provides an output that is much more readily interpreted than the intensity curve in terms of specific information on the protein structure.

In principle, the intensity distribution, $I(q)$, and the pair distribution function, $P(r)$, are related to each other by the pair of transforms



$$I(q) = 4\pi \int_0^{d\,\max} \frac{P(r)\sin(qr)}{qr}\,\mathrm{d}r \qquad (1)$$

and

$$P(r) = r/2\pi^2 \int_0^{\infty} I(q)\, q \sin(qr)\,\mathrm{d}q. \qquad (2)$$

The predicted scattering, $I(q)$, may be readily obtained for a given $P(r)$ as the extent of the $P(r)$ is bounded by the maximum dimension of the molecule, $d_{\max}$ [equation (1)]. However, the calculation of the inverse transform, to obtain $P(r)$ from $I(q)$ [equation (2)], is more problematic since an experimental data set is necessarily incomplete. Data may only be measured to a minimum value of q due to the finite size of the beam stop that shields the detector from the direct beam, and to a maximum value of q at the edge of the detector. The absence of these data from the calculation corrupts the appearance of $P(r)$, typically resulting in ripples running through the output function. It has also been suggested (Svergun, 1992) that when experimental intensities are used in direct calculations the associated errors may cause excessive inaccuracies in the resulting $P(r)$.

For these reasons the $P(r)$ calculated using SAXS data obtained from protein solution samples are invariably obtained by 'indirect' transform methods (Glatter, 1977; Moore, 1980; Svergun *et al.*, 1988). More recent work has continued to investigate this approach (Hansen & Pedersen, 1991; Liu & Zwart, 2012; Grant, 2022). All accounts that describe practical calculations of $P(r)$ (for example, in tutorial presentations and software documentation) consider the use of the indirect approach mandatory. Published work does not appear to document any examples of the successful use of direct transforms for the calculation of $P(r)$ in SAXS applications involving protein solution samples. A program intended for direct transform calculations is briefly noted in a paper describing updates to the *ATSAS* software suite (Manalastas-Cantos *et al.*, 2021) but no results are described, and use of this program has not been encouraged.

With the indirect transform approach to the calculation of $P(r)$, a set of functions is used to model $P(r)$ and the transform of these functions is fitted to $I(q)$. The resulting $P(r)$ combines the requirement that the predicted $I(q)$ agrees with the intensity data to within expected error and necessary physical attributes of the function including positivity and smoothness are enforced. In principle, there is a balance between the level of agreement of the predicted intensities with the data and these expectations for $P(r)$ (Svergun, 1992), and there exists a family of acceptable solutions that vary depending on how these aspects are weighted. In addition, the indirect transform method requires the assignment of the unknown structural parameter $d_{\max}$ as an *input* rather than $d_{\max}$ emerging as a structural *result*. The value of $d_{\max}$ may be assigned by visual inspection, by matching a set of heuristic 'perceptual criteria' (Svergun, 1992) or through a statistical analysis of $P(r)$ solutions (Hansen, 2000).

Indirect transform methods have undoubtedly proven to be very successful in practice, but these considerations motivate a proper examination of the calculation of $P(r)$ using direct transform methods. The naïve approach to the calculation, in which significant low-resolution data are missing and the measured intensities end abruptly at limited resolution, is inherently insufficient. However, commonly used methods in image reconstruction, involving the substitution of calculated data for missing data points at low resolution and smooth truncation of the data at the high-resolution edge, may be applied to ameliorate this situation. This paper describes methods and results of direct calculations of $P(r)$ using these techniques.

2. Materials and methods

2.1. Extending the measured intensity curve

Several steps were taken to extend and regularize the measured intensity data before use as the input for a direct transform.

Low-resolution data absent from the experimental measurements were interpolated using the Guinier equation

$$I(q) = I(0)\exp(-q^2R_g^2/3). \qquad (3)$$

The values for the intensity at zero scattering angle, $I(0)$, and the radius of gyration, R_g, were obtained by fitting a Guinier curve to measured data in the resolution range $\{0.65/\mathrm{Rg_{input}}, 1.3/\mathrm{Rg_{input}}\}$ (Grant *et al.*, 2015), where $\mathrm{Rg_{input}}$ is an initial input estimate for R_g. For unusual cases (for example, highly elongated proteins), an option is also available for the user to input a different chosen value for the upper resolution limit of the Guinier region.

The interval between extrapolated data points was set equal to the average spacing between measured data points. To ensure exact continuity and regularity for both extrapolated and measured data points, the experimental data were interpolated onto a new grid with this spacing.

Beyond the resolution limit of the measured data, $q_{\max}$, intensities were extrapolated to higher resolution using a q^{-4} decay function (Porod's law) to the point where the predicted intensity was 10% of the intensity at $q_{\max}$. To avoid problems arising from a particularly aberrant intensity measurement at $q_{\max}$ (where the data are typically noisy), the starting value for the extrapolated intensity was taken as the average of the five outermost data points. A constant background scattering intensity equal to the predicted intensity at the resolution limit of the extrapolated intensities was then subtracted from the complete modified intensity function so that it terminates at zero intensity.

In most practical cases, SAXS data will have been measured to a resolution beyond where there are significant modulations in the intensity curve and the intensity at $q_{\max}$ will be the approximate minimum for the data set. To manage occasional non-ideal cases, where the intensity at $q_{\max}$ is significantly higher than an intensity minimum at lower resolution (*cf.* data set SASDD42, Section 3.2), the modified intensity curve is

further processed by applying a fourth-order Butterworth filter, $B(q) = 1/\sqrt{[1+ (q/q_{\rm thresh})^4]}$ (Butterworth, 1930), where $q_{\rm thresh}$ is set to the resolution where the intensity is a minimum. A Butterworth filter retains the relative values of the input intensities to a resolution of $q_{\rm thresh}$ as faithfully as possible and then rapidly and smoothly reduces the intensity values at higher resolution.

This background-subtracted intensity function, with interpolated data at low q, regularized experimental data and extrapolated data beyond the experimentally measured $q_{\rm max}$ were used as the input for the direct calculation of $P(r)$.

To provide estimates for errors in $P(r)$ resulting from errors in the intensity data, $\sigma(I)$, the following procedure was used. Random values taken from Gaussian distributions with width $\sigma(I)$ were chosen at each data point, q. This set of values was used as the input for a direct transform so as to propagate them into a real-space function comparable to $P(r)$. To sample a representative range of possible error values from the given $\sigma(I)$ this process was repeated over 20 trials. For each point, r, in this set of functions the r.m.s. value was calculated and taken as a measure of the error in $P(r)$.

2.2. Display of results and estimation of $d_{\rm max}$

The values for $P(r)$ extending to $r = 4.5R_{\rm g}$ were calculated by direct transform from the extrapolated and modified intensity values. This range is somewhat greater than the largest physically plausible $d_{\rm max}$ for a given $R_{\rm g}$. Plots of $I(q)$ and $P(r)$ were displayed for visual review using the *PRIMUS* program from the *ATSAS 3.2.1* software suite. Ideally, the value of $P(r)$ is zero beyond $d_{\rm max}$ but noise in the data affects the solution in various ways. Displaying $P(r)$ over an extended range provides insight into the nature of the noise and is helpful for making a visual assessment of an appropriate value for $d_{\rm max}$.

For comparative purposes and to provide consistency with calculations of $P(r)$ by indirect transform programs, an additional output file, in which all values of $P(r)$ for $r > d_{\rm max}$ are truncated to zero, is also generated. To ensure a natural, smooth termination of $P(r)$ to zero value at $d_{\rm max}$, a linear reduction in the values of points in the last 10% of the range of $P(r)$ is applied if the function values do not decrease rapidly enough. The resulting $P(r)$ are normalized to a preset value using the area under the curve from zero to $d_{\rm max}$.

A method was also implemented to automatically estimate a minimum likely value for $d_{\rm max}$. A fourth-order Butterworth filter for error reduction (smoothing) was applied to the modified intensity data extending to $q_{\rm max}$. The value of $q_{\rm thresh}$ in this filter was set to min{0.2, $q_{\rm max}$/2} for q measured in Å^{-1}. The value of $d_{\rm max}$ was then estimated to be slightly larger (5% was used) than the value of r where this smoothed pair distribution function falls to 4% of the peak value, *i.e.* at a point where the smoothed $P(r)$ is losing significance.

2.3. Evaluation of $P(r)$ obtained by direct transform

In an initial set of tests, the accuracy of the direct transform method for calculating $P(r)$ was evaluated by establishing several reference protein-like $P(r)$ distributions. Atomic coordinates for structures that represent a variety of molecular sizes and shapes (Badger, 2019) were obtained from the Protein Data Bank (PDB; Berman *et al.*, 2000) (Fig. 1). Ideal target $P(r)$ were established by calculating histograms of atomic pair distances with bin widths of 1 Å from these structures. To represent some ambient fluctuation in the atomic positions, the value of each bin was smoothed slightly by mixing the value of each point, i, in $P(r)$ with neighboring values such that $P(r)_i = 0.25P(r)_{i-1} + 0.5P(r)_i + 0.25P(r)_{i+1}$. Synthetic data sets, $I(q)$, were calculated from these 1D distributions (Fig. 2). Data below a low-resolution limit, $q_{\rm min}$ = 0.0154 Å^{-1}, were rejected to mimic the typical absence of SAXS data at very low resolution and data were truncated at several different values of $q_{\rm max}$ (Table 1). The direct transform method described here was used to calculate $P(r)$ from these synthetic data sets and they were compared with the ideal $P(r)$ generated directly from the models (Fig. 3).

In a second set of tests, models (Fig. 4) and SAXS data (Fig. 5) were obtained from the Small Angle Scattering Biological Data Bank (SASBDB) (Kikhney *et al.*, 2020). To avoid overt

Table 1
Structural models used to calculate synthetic data to test the calculation of $P(r)$ by the direct transform.

For each example (PDB ID) the number of amino acids (No. AA) and the value of $d_{\rm max}$ are given. The ratio of $I(q)$ at $q_{\rm max}$ to $I(0)$ is given for $q_{\rm max}$ set to 0.5, 0.3, 0.2 and 0.15 Å^{-1}. The value of $\Delta P(r)$ reports the relative error in the $P(r)$ computed from the synthetic data with respect to the exact $P(r)$ obtained from a histogram of atom pairs in the model (see main text for details). Results for $\Delta P(r)$ are given for $q_{\rm max}$ set to 0.5, 0.3, 0.2 and 0.15 Å^{-1}. $I(q)$ and $P(r)$ obtained from entry 8r2w contain monomeric and dimeric forms, and $I(q)$ and $P(r)$ for entry 1mux contain an ensemble of 30 unique conformations (see main text for details).

PDB ID	No. AA	$I(q)/I(0)$ (%)	$\Delta P(r)$ (%)
4wkg	3960	0.006, 0.027, 0.141, 0.520	0.17, 0.20, 0.45, 3.97
1ss8	3780	0.010, 0.028, 0.145, 0.428	0.24, 0.28, 0.43, 1.86
2fo0	465	0.092, 0.199, 1.142, 2.918	0.34, 0.59, 1.22, 1.89
1y57	452	0.079, 0.223, 1.315, 4.330	0.46, 0.64, 1.10, 2.54
5tar	332	0.112, 0.356, 1.550, 8.231	0.41, 0.96, 2.93, 6.51
253l	298	0.204, 0.623, 3.962, 13.967	0.75, 1.52, 2.97, 3.28
8r2w	290, 580	0.100, 0.405, 1.186, 4.861	0.49, 0.75, 2.06, 4.00
1mux	148 × 30	0.333, 1.169, 6.030, 13.128	0.90, 2.65, 4.03, 15.14

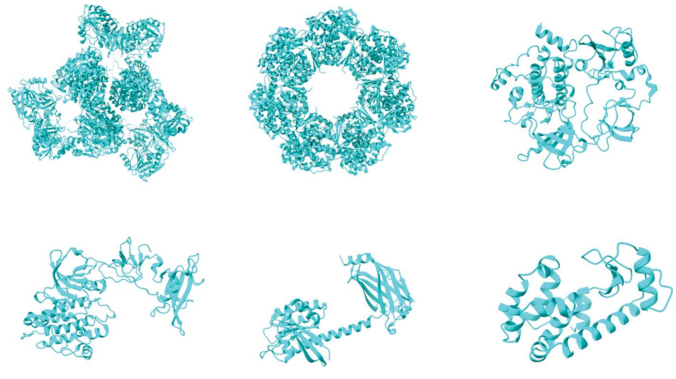


Figure 1
A set of structure models chosen to represent a variety of protein shapes and sizes. PDB IDs top row: 4wkg, 1ss8, 2fo0; bottom row: 1y57, 5tar, 253l. Molecular images were rendered with *ChimeraX* (Goddard *et al.*, 2018).

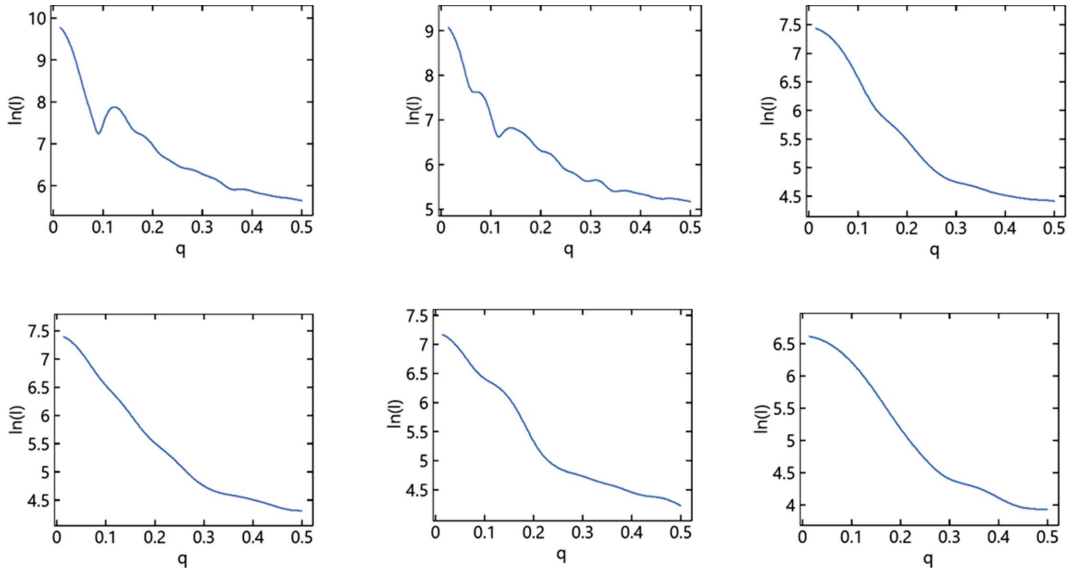


Figure 2
X-ray scattering intensities, $I(q)$, in the range $0.0154 < q < 0.5$ Å^{-1} computed from the pair distribution functions, $P(r)$, associated with the set of models shown in Fig. 1. Entries are ordered as in Fig. 1.

bias in example selection, these data sets correspond to a set of entries deposited within a specific time frame for which both plausible atomic models and indirect transform calculations with *GNOM* are available (Badger, 2019). The experimental data were used as the input for this direct transform methodology, and the resulting $P(r)$ were compared with the $P(r)$ that had been previously obtained with the indirect transform calculation program *GNOM* (Fig. 6) and deposited in the SASBDB.

2.4. Software availability

A simple Python script, *pyPr*, was used to perform the calculations described here. The script is available from the web site https://saxs2shapes.com and may be freely obtained under the GNU GPLv3 software license. For the calculations described here, the code was run on the Windows 11 operating system on a Dell Inspiron 16 Plus using the Python 3.12 interpreter. It is anticipated that this script will run on any computer hardware and operating system on which a Python interpreter is installed. Typical run times are sub-second.

3. Results and discussion

3.1. Performance of the direct transform for recovering model $P(r)$ functions

Calculations that compare the exactly known model $P(r)$ and the $P(r)$ obtained from the associated synthetic data, limited to ranges typically measured in SAXS studies, establish the inherent accuracy of the direct transform approach.

Table 1 shows the percentage error in $\Delta P(r)$ when the $P(r)$ calculated by the direct transform is compared with the exact $P(r)$ over the range $0 < r < d_{\max}$ [here, $d_{\max}$ is known from the

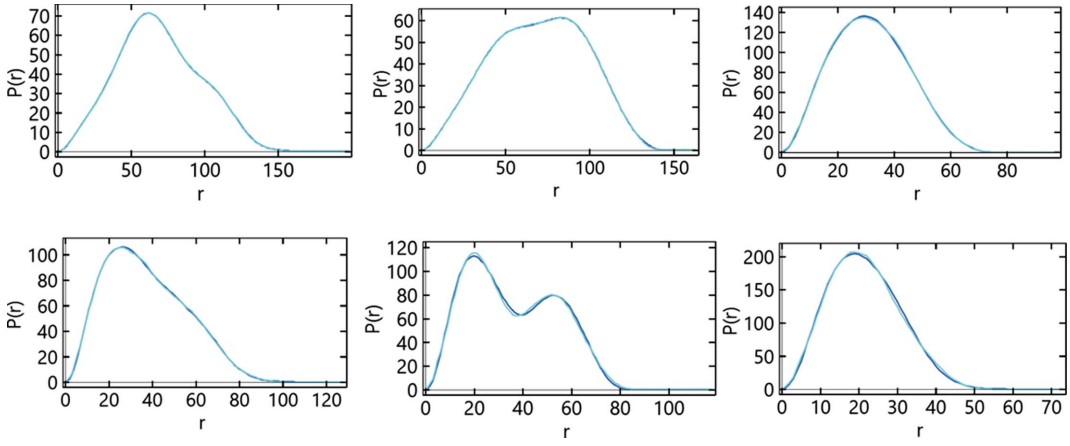


Figure 3
Pair distribution functions, $P(r)$, computed by the direct transform (dark blue curves) using simulated intensities to $q_{\max} = 0.2$ Å^{-1}. Pair distances, r, are measured in Å units. The $P(r)$ plots obtained directly from the atomic models (light blue curves) are almost coincident with the plots obtained by direct transform from the simulated intensity data. Entries are ordered as in Figs. 1 and 2.

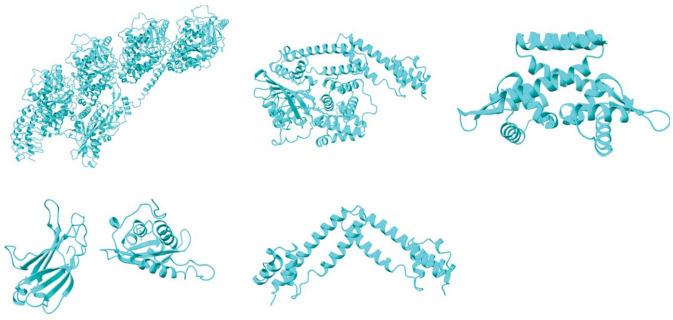


Figure 4
Atomic models associated with experimental X-ray solution scattering data obtained from the SASBDB. Top row: SASDCR9, SASDDG9, SASDD89; bottom row: SASDD76, SASDD42. Only atoms corresponding to crystallographically determined coordinates are rendered. Molecular images were rendered with *ChimeraX* (Goddard *et al.*, 2018).

atomic models used to build the $P(r)$]. The synthetic data associated with PDB entries 4wkg (Fischer *et al.*, 2015), 1ss8 (Chaudhry *et al.*, 2004), 2fo0 (Nagar *et al.*, 2006), 1y57 (Cowan-Jacob *et al.*, 2005), 5tar (Dharmaiah *et al.*, 2016) and 253l (Shoichet *et al.*, 1995) correspond to single conformations and are intended to typify the types of scattering patterns obtained from relatively stable, well ordered proteins. With the limit $q_{max} = 0.3$ Å^{-1} the difference in the calculated $P(r)$ compared with the exact $P(r)$ is always less than 1.52%, and even when the data were limited to $q_{max} = 0.2$ Å^{-1} the error only increased to 3% for the two smallest structures. Plots of the exact and calculated $P(r)$ functions (Fig. 3) for $q_{max} = 0.2$ Å^{-1} are almost indistinguishable, and the calculated $P(r)$ do not show discernible ripples or artifacts. Tabulations of $I(q_{max})/I(0)$ demonstrate that in all these examples the scattered intensities are less than 1% of $I(0)$ at a resolution of $q_{max} = 0.3$ Å^{-1}. It appears, therefore, that the significant data needed for the calculation of $P(r)$ are contained within a moderate resolution limit. SAXS intensities fall the most rapidly for the larger structures and the $I(q_{max})/I(0)$ ratios are the smallest, leading to the most accurate calculations of $P(r)$. Even in the most unfavorable cases, provided $q_{max} > 0.2$ Å^{-1}, the calculation errors inherent in this formulation of the direct transform method appear similar to or less than the expected experimental measurement errors of 1–3%. At least with error-free data, it appears that the direct transform method formulated here can accurately calculate the pair distribution function for these types of protein samples.

When q_{max} was reduced to 0.15 Å^{-1} the algorithmic inaccuracy increased and the error in the calculated $P(r)$ was in the range 1.9–3.3% for 1ss8, 2fo0, 253l and 1y57. More visible distortions were evident for 4wkg (error of 4%). These calculated $P(r)$ would still be usable for semi-quantitative work, but the errors might be considered somewhat large for a very precise analysis. The $P(r)$ calculated for 5tar (error of 6.5%) was clearly deformed, with much less distinction in the two maxima than in the reference distribution. The largest inaccuracies in the calculated $P(r)$ occur when the missing sections of higher-resolution data cover a large intensity range and contain features that are not adequately approximated by the Porod law (q^{-4}) extrapolation. However, standard data collection procedures should not normally truncate the experimental data as severely as in these calculations. Measuring data out to a q_{max} that contains all significant features in the intensity curve should give the most reliable results.

3.2. Comparison of *P*(*r*) obtained by direct and indirect transforms with experimental data

To investigate the performance of the direct transform method with experimental data, calculations were performed on five examples taken from the SASBDB and compared with the associated $P(r)$ obtained with the program *GNOM*

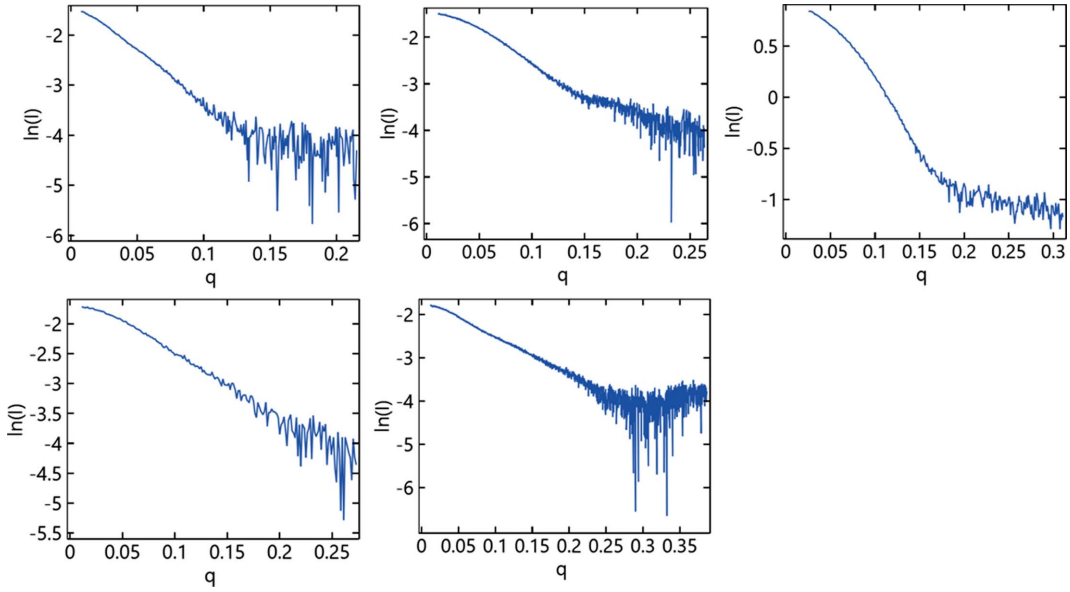


Figure 5
Experimentally determined X-ray scattering intensities, $I(q)$, for the SASBDB IDs associated with the set of models shown in Fig. 4. Entries are ordered as in Fig. 4.

Table 2
Data sets obtained from the SASBDB used to test the calculation of $P(r)$ by direct transform.

For each example (SASBDB ID) the number of amino acids (No. AA) is given for the associated model. The range of available data (q_{min}, q_{max}), the ratio of $I(q)$ at q_{max} to $I(0)$ and the published value of d_{max} (pub) are given for each entry. A minimal estimate for d_{max} (est), estimated by the software, is listed. The R factor compares the experimental data with the back-transform of the direct calculation of $P(r)$ when terminated at d_{max}(pub). The value of $\Delta P(r)$ reports the relative difference of the $P(r)$ obtained by the direct transform with respect to the $P(r)$ obtained using the indirect transform program *GNOM*, deposited for each entry.

SASBDB ID	No. AA	q_{min}, q_{max} (Å^{-1})	$I(q)/I(0)$ (%)	d_{max}(pub, est) (Å)	R (%)	$\Delta P(r)$ (%)
SASDCR9	2418	0.009, 0.215	0.129	195, 184	2.02	0.89
SASDDG9	564	0.011, 0.264	0.214	130, 98	1.68	1.48
SASDD89	263	0.027, 0.311	0.895	75, 67	1.43	2.72
SASDD76	297	0.013, 0.273	0.335	93, 84	2.30	3.88
SASDD42	209	0.013, 0.387	0.897	105, 94	2.30	2.24

(Table 2, Fig. 4). The R factors were obtained from the reverse transform of the $P(r)$ obtained by the direct transform method after termination at d_{max} and mainly reflect the point-to-point 'jitter' (random error) in the experimental data. Overall, both direct and indirect transform methods give similar results for $P(r)$ (Fig. 6) and the differences between them (0.9–3.9%) are comparable to the expected error in the data.

A somewhat troubling aspect of this study is that the published values for d_{max} frequently result in $P(r)$ that appear to extend beyond the more compact shapes typical of protein structures. These tails in $P(r)$ appear in calculations performed by both direct and indirect transform methods, so they would appear to be an inherent aspect of the experimental data and not an artifact of the specific calculation method. In examples SASDCR9 (Trofimova *et al.*, 2018) and SASDDG9 (Sluchanko *et al.*, 2018) visual assessments and considerations of the expected structures suggest that the $P(r)$ functions might be expected to terminate more naturally at $d_{max} \sim$ 180 Å and $d_{max} \sim$ 100 Å, respectively. SASDDG9 represents an extreme case, where the published value of d_{max} is more than four times the R_g of ∼32 Å and beyond expected limits. However, in both cases the magnitude of $P(r)$ in the tail region is relatively small. For SASDD89 (Kutnowski *et al.*, 2018), SASDD76 (Chen *et al.*, 2018) and SASDD42 (Slonimskiy *et al.*, 2018) the tails are of larger magnitude and extend d_{max} more than 10% beyond ∼65, ∼80 and ∼90 Å, respectively, which visual estimates might suggest as more typically protein-like. The more minimal estimates for d_{max} produced by *pyPr* (Table 2) truncate these $P(r)$ at d_{max} close to these expectations. Choosing values for d_{max} that result in the most physically reasonable shape for $P(r)$ would eliminate these (possibly artifactual) tail features which may be unrelated to the underlying protein structure but would also conceal an aspect of the data from the reader. With the direct transform the entire $P(r)$, including the region beyond a preset d_{max}, is revealed, whereas with the indirect transform information on $P(r)$ beyond that point is voided.

3.3. Validity of data modeling

The Guinier law [equation (3)] was used to calculate and fill in missing intensity values at very low resolution, typically where $q_{min} < \sim 0.015$ Å^{-1}. The two constants, $I(0)$ and R_g, needed to model these intensities do not depend on knowledge of the atomic structure but are obtained by the *pyPr* software by fitting the observed data within the Guinier region. For a typical globular protein, for example, characterized by $R_g = 30$ Å, the Guinier region will usually extend to $q \sim 0.04$ Å^{-1} so there will generally be ample observed data in this region for an accurate evaluation. In a SAXS analysis the extent of the Guinier region may be defined operationally by finding the value of q where the plot of ln(I) versus q^2 becomes non-linear. For the calculations within the *pyPr* program this upper limit may be input directly on the basis of a prior Guinier analysis or set via the widely used relationship $1.3/Rg_{input}$, where Rg_{input} is based on the prior analysis. It is generally considered that Guinier's law *must* be obeyed for the lowest-resolution data for a satisfactory SAXS analysis of

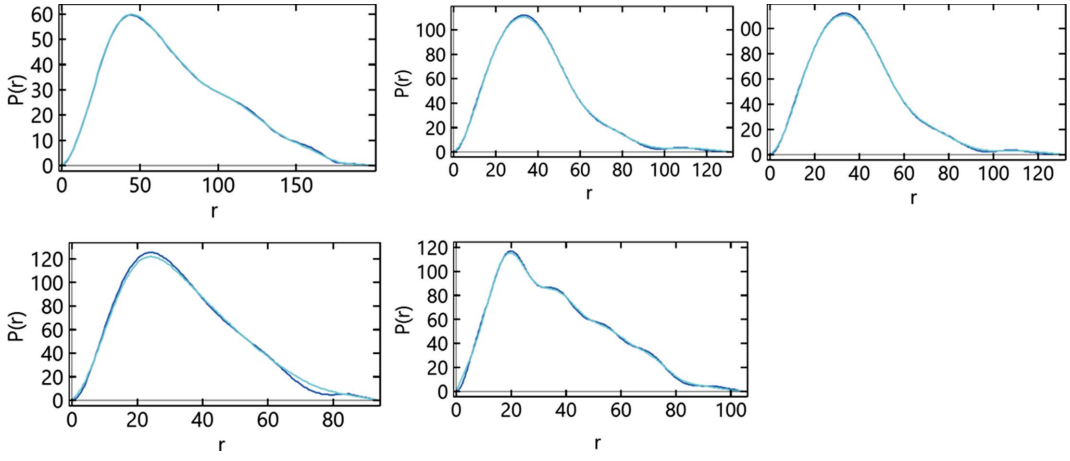


Figure 6
Comparison of pair distribution functions, $P(r)$, computed by the direct transform (dark blue curves) and the indirect transform using *GNOM* as deposited in the SASBDB (pale blue curves). Pair distances, r, are measured in Å units. The $P(r)$ are almost coincident over most of the range, which is terminated at the published values of d_{max}. Entries are ordered as in Figs. 4 and 5.

protein structure and represents an initial quality control on the utility of the protein sample (Trewhella *et al.*, 2017). (If this is not the case, due to interference effects between protein molecules or aggregation, the data set will be considered compromised for reliable analysis.) Thus, the Guinier law provides the proper basis for the calculation of missing low-q data for the types of samples and studies described in this work.

Intensities in the high-q regime, beyond the limit of the measured data, were extended using Porod's law, which posits a q^{-4} decay in intensity values, and this law is considered the appropriate approximation for the fall-off in high-q scattering from compact, well folded protein structures. The calculation of this intensity decay function does not contain or require any information on the protein structure. Ideally, SAXS data will have been measured to beyond the limit of a significant signal, to a point where the intensities are near zero, and the impact of this correction is, therefore, necessarily small. Some calculations using the series of synthetic data sets truncated at the point q_{max} = 0.2 Å^{-1}, where the input intensities are quite limited (Fig. 2), provide a sense of the magnitude and usefulness of the Porod law extrapolation. For the two largest structures, 4wkg and 1ss8, $I(q)/I(0)$ is ∼0.14% at q_{max} and the fit between the calculated and reference $P(r)$ would have been worsened by only ∼0.4% if the Porod law extrapolation had been omitted. For 2fo0, 1y57 and 5tar, $I(q)/I(0)$ is 1.1–1.6% at q_{max} and the extrapolation has a more significantly beneficial impact. In these cases, the agreement between the calculated and reference $P(r)$ cited in Table 1 would have been worsened by 3–5% if the Porod law extrapolation had not been applied. For the smallest structure, 253l, the Porod law extrapolation was the most impactful and the error in the calculation of $P(r)$ would be 10% higher if it were omitted. Thus, the use of the Porod law data extrapolation appears practically justified and helpful for calculations of $P(r)$ from intensity data obtained from samples that consist of compact proteins in solution.

With the direct transform method, the modeling of data missing from $I(q)$ is explicit and applies established expectations for intensity data from well ordered globular proteins (Guinier and Porod laws) to substitute for the missing intensities. This methodology contrasts with the indirect transform method, which does not explicitly model missing data but employs a well chosen set of adjustable functions to model $P(r)$. Implicit in the application of indirect transform methods is that the modeling functions are able to accurately represent the true $P(r)$ and that correct fits of the transforms of these functions to the data are achievable using only the observed $I(q)$.

3.4. Structural ensembles

The direct transform method described here was tested using synthetic and real intensity data obtained from well ordered globular proteins in dilute solutions. These types of samples are common in many biological SAXS experiments, for example to probe changes in protein conformation upon enzyme activation or to determine the structure of a protein complex when the structures of the separate components are known. For samples that contain proteins in multiple conformations, it might be expected that $I(q)$ would be somewhat smoother than in cases that involve only a single conformation and would decay at higher q more slowly than Porod's law

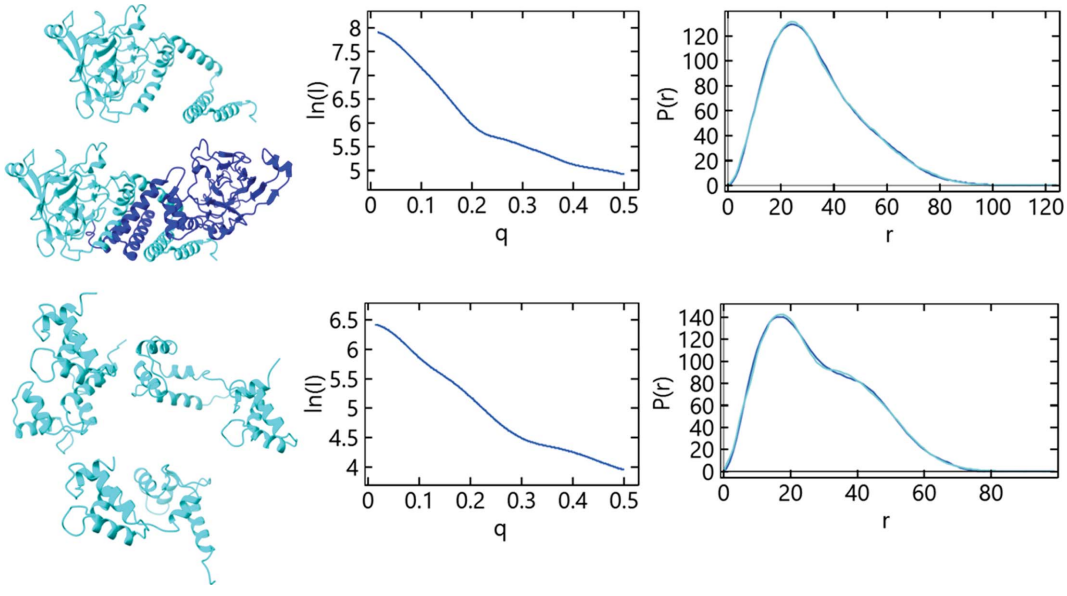


Figure 7
Atomic models, synthetic data and $P(r)$ for PDB entries 8r2w (top row) and 1mux (bottom row). For 8r2w the target $P(r)$ was constructed from a 4:1 mixture of monomers (pale blue) to dimers (pale and dark blue). For 1mux the target $P(r)$ was constructed from an ensemble of 30 structural forms (three examples are shown). Molecular images were rendered with *ChimeraX* (Goddard *et al.*, 2018). The middle panels show the X-ray scattering intensities, $I(q)$, in the range $0.0154 < q < 0.5$ Å^{-1} computed from the reference pair distribution functions, $P(r)$, associated with these models. The right panels show pair distribution functions, $P(r)$, computed by the direct transform (dark blue curves) using simulated intensities to q_{max} = 0.2 Å^{-1} for 8r2w and q_{max} = 0.3 Å^{-1} for 1mux. These computed distributions are compared with the reference $P(r)$ (pale blue). Pair distances, r, are measured in Å units.

would predict. The direct transform method has not been evaluated for the calculation of $P(r)$ using experimental data from proteins in mixed oligomeric forms or in highly disordered states, but the potential for success from these types of samples was evaluated using model data.

Tests for entry 8r2w (Kuatsjah *et al.*, 2024) evaluated the case where the protein sample contains both monomers and dimers. $P(r)$ and $I(q)$ were calculated for each form separately (*cf.* Section 2.3) and added together. Since the magnitudes of $P(r)$ and $I(q)$ depend on the number of atom pairs in the protein, the values for the monomeric form were weighted by a factor of four so that both structural forms made equal contributions to these functions. Although this protein is composed of both monomers and dimers, with different values for R_g and different ranges for Guinier and Porod regions, the resulting $P(r)$ calculated by the direct transform appears to be accurate (Table 1, Fig. 7).

Tests for entry 1mux (Osawa *et al.*, 1998) evaluated a case where the protein sample contains a diverse range of conformations. This protein was solved by NMR spectroscopy and so the PDB entry contains 30 models that cover a diverse range of conformations with d_{max} ranging from 57 to 89 Å. Here, reference $P(r)$ were constructed for each conformation, $I(q)$ was calculated from each example, and these sets of $P(r)$ and $I(q)$ were combined to create a composite target $P(r)$ and associated synthetic data set $I(q)$. The $P(r)$ obtained from these calculations correctly reproduces the target $P(r)$ when the input data extend to $q_{max} = 0.3$ Å^{-1} (Table 1, Fig. 7) but the accuracy of the results with more limited data is somewhat worse than for the other examples. When applying the direct transform method on structurally heterogeneous samples, extra care may be needed to collect experimental data that are as complete as possible.

4. Summary

The examples presented here challenge the universally stated assertion that the $P(r)$ obtained during a SAXS analysis of protein samples must be calculated using the indirect transform method to avoid artifacts and significant errors. Provided that appropriate steps are implemented to ameliorate the absence of data from the measured intensity spectrum, the direct transform method provides $P(r)$ that are very similar to those obtained with the most popular indirect transform algorithm, *GNOM*. The magnitudes of the experimental errors that occur in practice do not appear to significantly affect the quality of the $P(r)$ obtained by direct transform relative to the results obtained by indirect transform.

In test calculations with synthetic data the target $P(r)$ distributions were set up by a direct evaluation of atom pair distances from within various atomic models. These model $P(r)$ were constructed on fine (1 Å) grids which are, in principle, capable of representing very high resolution detail in pair distance distributions. However, the $P(r)$ associated with compact, globular protein structures are inherently smoothly varying and lack sharp features. This is because a protein containing a few hundred atoms will contain over 100000 atomic pairs and when these pair distances are projected into 1D histograms they form a smooth continuum extending to d_{max}. For this reason, the strength of the X-ray scattering declines rapidly with increasing q and intensity data to the resolution limits typically measured in SAXS experiments ($q \sim 0.3$ Å^{-1}) are able to accurately capture the information needed to calculate $P(r)$.

The calculation of $P(r)$ is a routine step in SAXS structure analysis but is not typically a major focus of the investigation. For example, it is uncommon to compare the experimentally determined $P(r)$ with the $P(r)$ predicted by a crystallographic atomic model, although such a comparison could help identify how solution and crystal conformations differ. Although very informative, popular 3D shape reconstruction methods (Svergun, 1999; Svergun *et al.*, 2001; Grant, 2018) for calculating the 3D envelopes for protein molecules from SAXS data are inherently limited to low resolution by the data modeling assumptions (Svergun *et al.*, 2001) and somewhat indefinite due to the multi-solution Monte Carlo methods used in the solution process. The reconstructed molecular envelopes are not accurate enough to form reliable quantitative conclusions about, for example, shifts of ∼5 Å in the relative distance between two domains. A $P(r)$ analysis, albeit restricted to one dimension, is only limited by the resolution of the data and will typically provide more accurate and quantifiable results on differences between protein structures.

Conflict of interest

There are no conflicts of interest.

Data availability

Atomic models were obtained from the Protein Data Bank and SAXS data were obtained from the Small Angle Scattering Biological Data Bank.

References

Badger, J. (2019). *J. Appl. Cryst.* **52**, 937–944.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Billinge, S. J. L. (2019). *Philos. Trans. R. Soc. A.* **377**, 20180413.

Butterworth, S. (1930). *Exp. Wireless Wireless Eng.* **7**, 536–541.

Chaudhry, C., Horwich, A. L., Brunger, A. T. & Adams, P. D. (2004). *J. Mol. Biol.* **342**, 229–245.

Chen, K.-E., Tillu, V. A., Chandra, M. & Collins, B. M. (2018). *Structure*, **26**, 1612–1625.e4.

Cowan-Jacob, S. W., Fendrich, G., Manley, P. W., Jahnke, W., Fabbro, D., Liebetanz, J. & Meyer, T. (2005). *Structure*, **13**, 861–871.

Dharmaiah, S., Bindu, L., Tran, T. H., Gillette, W. K., Frank, P. H., Ghirlando, R., Nissley, D. V., Esposito, D., McCormick, F., Stephen, A. G. & Simanshu, D. K. (2016). *Proc. Natl Acad. Sci. USA*, **113**, e6766.

Fischer, U., Hertlein, S. & Grimm, C. (2015). *Acta Cryst.* D**71**, 687–696.

Glatter, O. (1977). *J. Appl. Cryst.* **10**, 415–421.

Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H. & Ferrin, T. E. (2018). *Protein Sci.* **27**, 14–25.

Grant, T. D. (2018). *Nat. Methods*, **15**, 191–193.
Grant, T. D. (2022). *J. Appl. Cryst.* **55**, 1116–1124.
Grant, T. D., Luft, J. R., Carter, L. G., Matsui, T., Weiss, T. M., Martel, A. & Snell, E. H. (2015). *Acta Cryst.* D**71**, 45–56.
Hansen, S. (2000). *J. Appl. Cryst.* **33**, 1415–1421.
Hansen, S. & Pedersen, J. S. (1991). *J. Appl. Cryst.* **24**, 541–548.
Kikhney, A. G., Borges, C. R., Molodenskiy, D. S., Jeffries, C. M. & Svergun, D. (2020). *Protein Sci.* **29**, 66–75.
Kuatsjah, E., Schwartz, A., Zahn, M., Tornesakis, K., Kellermyer, Z. A., Ingraham, M. A., Woodworth, S. P., Ramirez, K. J., Cox, P. A., Pickford, A. R. & Salvachúa, D. (2024). *Cell. Rep.* **43**, 115002.
Kutnowski, N., Shmuely, H., Dahan, I., Shmulevich, F., Davidov, G., Shahar, A., Eichler, J., Zarivach, R. & Shaanan, B. (2018). *J. Struct. Biol.* **204**, 191–198.
Liu, H. & Zwart, P. H. (2012). *J. Struct. Biol.* **180**, 226–234.
Manalastas-Cantos, K., Konarev, P. V., Hajizadeh, N. R., Kikhney, A. G., Petoukhov, M. V., Molodenskiy, D. S., Panjkovich, A., Mertens, H. D. T., Gruzinov, A., Borges, C., Jeffries, C. M., Svergun, D. I. & Franke, D. (2021). *J. Appl. Cryst.* **54**, 343–355.
Moore, P. B. (1980). *J. Appl. Cryst.* **13**, 168–175.
Nagar, B., Hantschel, O., Seeliger, M., Davies, J. M., Weis, W. I., Superti-Furga, G. & Kuriyan, J. (2006). *Mol. Cell*, **21**, 787–798.
Osawa, M., Swindells, M. B., Tanikawa, J., Tanaka, T., Mase, T., Furuya, T. & Ikura, M. (1998). *J. Mol. Biol.* **276**, 165–176.
Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. (1995). *Proc. Natl Acad. Sci. USA*, **92**, 452–456.
Slonimskiy, Y. B., Maksimov, E. G., Lukashev, E. P., Moldenhauer, M., Jeffries, C. M., Svergun, D. I., Friedrich, T. & Sluchanko, N. N. (2018). *Biochim. Biophys. Acta*, **1859**, 382–393.
Sluchanko, N. N., Slonimskiy, Y. B., Shirshin, E. A., Moldenhauer, M., Friedrich, T. & Maksimov, E. G. (2018). *Nat. Commun.* **9**, 3869.
Svergun, D. I. (1992). *J. Appl. Cryst.* **25**, 495–503.
Svergun, D. I. (1999). *Biophys. J.* **76**, 2879–2886.
Svergun, D. I., Petoukhov, M. V. & Koch, M. H. J. (2001). *Biophys. J.* **80**, 2946–2953.
Svergun, D. I., Semenyuk, A. V. & Feigin, L. A. (1988). *Acta Cryst.* A**44**, 244–250.
Trewhella, J., Duff, A. P., Durand, D., Gabel, F., Guss, J. M., Hendrickson, W. A., Hura, G. L., Jacques, D. A., Kirby, N. M., Kwan, A. H., Pérez, J., Pollack, L., Ryan, T. M., Sali, A., Schneidman-Duhovny, D., Schwede, T., Svergun, D. I., Sugiyama, M., Tainer, J. A., Vachette, P., Westbrook, J. & Whitten, A. E. (2017). *Acta Cryst.* D**73**, 710–728.
Trofimova, D., Paydar, M., Zara, A., Talje, L., Kwok, B. H. & Allingham, J. S. (2018). *Nat. Commun.* **9**, 2628.