Journal of Structural Biology 180 (2012) 226–234

Contents lists available at SciVerse ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Determining pair distance distribution function from SAXS data using parametric functionals

Haiguang Liu [a,b,*], Peter H. Zwart [a,*]

[a] Physical Biosciences Division, Lawrence Berkeley National Laboratories, One Cyclotron Road, Berkeley, CA 94720, USA
[b] Arizona State University, Department of Physics, P.O. Box 871504, Tempe, AZ 85287-1504, USA

ARTICLE INFO

Article history:
Received 26 March 2012
Received in revised form 10 May 2012
Accepted 16 May 2012
Available online 1 June 2012

Keywords:
Pair distance distribution function
Small angle scattering
Parametric
Open source

ABSTRACT

Small angle X-ray scattering (SAXS) experiments are widely applied in structural biology. The SAXS experiments yield one-dimensional profile that needs further analysis to reveal structural information. The pair distance distribution function (PDDF), $P(r)$, can provide molecular structures more intuitively, and it can be used to guide *ab initio* model reconstructions, making it a critical step to derive $P(r)$ from experimental SAXS profiles. To calculate the $P(r)$ curves, a new method based on a specially designed parametric functional form is developed, and implemented in *pregxs*. This method is tested against both synthetic and experimental data, the estimated $P(r)$ functions are in good agreement with correct or known $P(r)$. The method can also predict the molecular size. In summary, the *pregxs* method is robust and accurate in $P(r)$ determination from SAXS profiles. The *pregxs* source code and an online server are available at http://www.sastbx.als.lbl.gov.

Published by Elsevier Inc.

1. Introduction

The knowledge of macromolecule structures is important to understand the molecular interactions and mechanism (Orengo et al., 1999). High resolution structures provide detailed information, but applications hinge on experimental limits: X-ray crystallography relies on the growth of good crystals that can survive high dose of X-rays during the data collection, while the nuclear magnetic resonance (NMR) method is mostly applicable to relative small molecules (Putnam et al., 2007; Madl et al., 2011). Furthermore, probing high resolution structural information often introduces additional constraints, such as crystal packing, thus making it more challenging to study the dynamics of molecules. The small angle X-ray scattering (SAXS) is an alternative technique used to study macromolecular structure and dynamics (Glatter and Kratky, 1982; Koch et al., 2003; Hura et al., 2009; Mertens and Svergun, 2010). Although the SAXS experiments only provide low resolution information, wider range structural and dynamics information can be probed by changing the buffer compositions. On the other hand, since the molecules are randomly oriented in solution, the scattering profile is the angular average of scattering patterns. The SAXS profiles are 1D curves ($I(q)$), where the intensity is the function of scattering angles. The scattering angle is usually converted to momentum transfer q ($q = 4\pi\sin(\theta)/\lambda$ where λ is the wavelength and 2θ is the scattering angle).

The scattering intensity profiles do not reveal intuitive structural information, which usually require some proper transformations of the SAXS profile, $I(q)$. For example, molecular weight and the radius of gyration (R_g) can be obtained from the Guinier analysis in the *Guinier* region, typically requiring ($0 \leqslant qR_g \leqslant 1.3$) for globular proteins (Koch et al., 2003). More structural information can be learned by determining the pair distance distribution function, or $P(r)$. The $P(r)$ function reveals valuable information about the shape of molecules, allowing more intuitive interpretation of the intensity profile, $I(q)$. Further more, the $P(r)$ is also critical in real space 3D model constructions. For example, the DAMMIF method requires accurately determined $P(r)$ to obtain dummy atom models that correspond to the scattering profiles (Franke and Svergun, 2009).

The $I(q)$ and $P(r)$ are closely related. It is straightforward to calculate the intensity profile, where

$$I(q) = \int_0^{d_{\max}} P(r)\frac{\sin(qr)}{qr}dr \tag{1}$$

which can be derived from the Debye formula (Debye, 1915), and the inverse transformation allows one to calculate $P(r)$:

$$P(r) = \frac{r}{2\pi^2}\int_0^{\infty} qI(q)\sin(qr)dq \tag{2}$$

Although this formula provides a direct way of calculating $P(r)$, it is not practical in reality, because the $I(q)$ is only available at

* Corresponding authors at: Physical Biosciences Division, Lawrence Berkeley National Laboratories, One Cyclotron Road, Berkeley, CA 94720, USA (H. Liu).
E-mail addresses: haiguang.liu@asu.edu, hgliu@lbl.gov (H. Liu), phzwart@lbl.gov (P.H. Zwart).

1047-8477/$ - see front matter Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.jsb.2012.05.011

discrete points for limited q-range. Noise introduced during the data collection and processing makes it more challenging to calculate $P(r)$ using the direct Fourier transformation (Koch et al., 2003).

Another approach to estimate $P(r)$ is by estimating $P(r)$ under proper assumptions and using Eq. (1) to calculate the intensity profile, then iteratively find the $P(r)$ that yields the optimal fit to the $I(q)$ profile. Using this approach, the estimation of $P(r)$ is converted to a constraint optimization problem: finding the $P(r)$ with associated $I(q)$ that agrees to experimental data while ensuring the satisfaction of $P(r)$ to the imposed constraints. Several methods have been implemented following this approach (Glatter, 1977; Hansen and Pedersen, 1991; Svergun, 1992; Müller and Hansen, 1994; Krauthauser and Nimtz, 1996; Hansen, 2000; Swain et al., 2001; Ilavsky and Jemian, 2009). The $P(r)$ is typically represented using a set of kernel functions. For example, Glatter proposed the usage of cubic B splines as the smooth kernel functions (Glatter, 1977), while Moore used a set of *sine* functions with different frequencies (Moore, 1980). The optimization goal is to minimize the χ^2 between the model intensity and the experimental data, while enforcing smoothness and non-negativeness of $P(r)$. Svergun and coworkers successfully developed $P(r)$ calculation routines with perceptual criterion evaluations to the yielded $P(r)$ (Svergun, 1992). To obtain the correct $P(r)$, the molecular size measured as the maximum distance between atoms, the $d_{\max}$, is required and usually has to be determined as *a priori*. Based on the radius of gyration, R_g, the $d_{\max}$ can be narrowed down to a specific range for globular proteins. Different values of $d_{\max}$ in the previously determined range will be tried to calculate the $P(r)$ associated with each $d_{\max}$. Then the $P(r)$ distributions should be visually inspected and figure out the optimal $P(r)$ together with the $d_{\max}$. Encouragingly, there is some recent progresses in reducing human interventions, for example, the program *AUTOGNOM* is capable of finding the optimal $d_{\max}$ (Petoukhov et al., 2007).

In this paper, we propose a method that determines the $P(r)$ using a parametric functional form with built-in smoothness and non-negative characteristics. Because this parametric functional form has intrinsic properties that match the $P(r)$ distributions of compact molecules, constraints are reduced. The method, *pregxs* (abbreviation for $P(r)$ estimation given X-ray scattering data), runs in two modes: optimize the $P(r)$ with the *prior* $d_{\max}$, or search both the $P(r)$ and the $d_{\max}$. This method was tested using both synthetic data calculated from PDB models and experimental data. The comparisons to the $P(r)$ derived from PDB models or calculated using other methods demonstrate that the *pregxs* obtains correct $P(r)$ from intensity profile $I(q)$.

2. Methods

In this section, the detailed method of using the specially designed parametric function to estimate the $P(r)$ is described. An efficient way of calculating intensity profiles is also derived, followed by SAXS profile comparison methodology and how to utilize *prior* knowledge to improve $P(r)$ estimation.

2.1. Parameterization

A probability density function (pdf) $p(x)$ with $(-1 \leqslant x \leqslant 1)$ can be expressed as the product of a *prior* pdf $g(x)$ and an exponentiated Chebyshev polynomial series:

$$p(x) = g(x)\exp\left[\sum_{m=0}^{M} a_m T_m(x)\right] \tag{3}$$

The *prior* pdf $g(x)$ is chosen to satisfy boundary conditions, i.e., $p(x)$ is zero at end points. The pair distance distribution function ($P(r)$) of unit sphere is the default *prior* pdf, $g(x)$ (note that $x = r - 1$ to scale the function to [-1,1], where r is the distance between atoms) (Glatter and Kratky, 1982). The positive exponential modification terms, together with the *prior* pdf, guarantee that the resulted function $p(x)$ is continuously non-negative throughout $[-1,1]$. While the *prior* pdf $g(x)$ sets the basic shapes for the $p(x)$, the coefficients of Chebyshev polynomials ($T_m(x)$) for the exponential modifier can be selected to sample a wide variety of probability distribution functions.

The Chebyshev polynomials are chosen as the modifier terms, because they are orthogonal in range $[-1,1]$ with weight $\frac{1}{\sqrt{1-x^2}}$ (Schwerdt, 1966). The Chebyshev polynomial at degree m has m roots, the resulted $p(x)$ can model curves with multiple modals. Due to the orthogonality of Chebyshev polynomials, more details of the $p(x)$ can be added by including higher order terms (Fig. 1a). By varying the coefficients $\{a_m\}$, the parametrized function $p(x)$ can also be changed, as demonstrated by Fig. 1b.

2.2. Intensity profile calculation

The $p(x)$ can be transformed into a (normalized) pair distance distribution function $P(r)$ via

$$r = \frac{d_{\max}}{2}(x+1) \tag{4}$$

with the associated Jacobian

$$J(r,x) = \frac{\partial r}{\partial x} = \frac{d_{\max}}{2} \tag{5}$$

For any given $P(r)$, the corresponding scattering profile can be computed using Eq. (1). Since the $P(r)$ is changed during the optimization process, the direct implementation of Eq. (1) is slow. To speed up the intensity calculation without losing much accuracy, the equation is rewritten as:

$$I(q) = \sum_i P(r_i)\int_{r_i}^{r_{i+1}} \frac{\sin(qr)}{qr}\,dr \tag{6}$$

setting

$$S(r_i,q) = \int_{r_i}^{r_{i+1}} \frac{\sin(qr)}{qr}\,dr \tag{7}$$

One gets a fast intensity calculation for any given $P(r)$, with $S(r_i,q)$ precomputed:

$$I(q) = \sum_i P(r_i)S(r_i,q) \tag{8}$$

2.3. Model-data correspondence

The pair distance distribution function $P(r)$ can be adjusted by changing the parameters $\{a_m\}$. The discrepancy between the observed data and the $P(r)$ associated intensity is measured using the χ^2 scoring function, defined as:

$$\chi^2 = \frac{1}{N_{\text{obs}}}\sum_{j=1}^{N_{\text{obs}}}\left[\frac{I_{\text{obs}}(q_j) - kI_{\text{calc}}(q_j)}{\sigma_j}\right]^2 \tag{9}$$

where the factor k is a scaling factor that is absorbed in the coefficient a_0.

2.4. Prior knowledge

When adjusting the coefficients $\{a_m\}$ in $P(r)$ to maximize the correspondence between calculated and observed scattering data (via Eq. (6)), one is at the risk of over-fitting the data when the number of parameters is large. Specifically, the inclusion of a large number of Chebyshev polynomials can introduce large oscillations