Special Issue
"Progress of Theoretical and Computational Biophysics"

Vol. 16, pp. 430–443 (2019)
doi: 10.2142/biophysico.16.0_430

Biophysics and Physicobiology
https://www.jstage.jst.go.jp/browse/biophysico/

Regular Article

Characterization of X-ray diffraction intensity function from a biological molecule for single particle imaging

Atsushi Tokuhisa[1,2,3]

[1] *RIKEN Cluster for Science and Technology Hub, Kobe, Hyogo 650-0047, Japan*
[2] *RIKEN Center for Computational Science, Kobe, Hyogo 650-0047, Japan*
[3] *RIKEN Medical Sciences Innovation Hub Program, Yokohama, Kanagawa 230-0045, Japan*

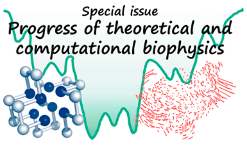
Received July 1, 2019; accepted September 3, 2019

An attainable structural resolution of single particle imaging is determined by the characteristics of X-ray diffraction intensity, which depend on the incident X-ray intensity density and molecule size. To estimate the attainable structural resolution even for molecules whose coordinates are unknown, this research aimed to clarify how these characteristics of X-ray diffraction intensity are determined from the structure of a molecule. The functional characteristics of X-ray diffraction intensity of a single biomolecule were theoretically and computationally evaluated. The wavenumber dependence of the average diffraction intensity on a sphere of constant wavenumber was observable by small-angle X-ray solution scattering. An excellent approximation was obtained, in which this quantity was expressed by an integral transform of the product of the external molecular shape and a universal function related to its atom packing. A standard model protein was defined by an analytical form of the first factor characterized by molecular volume and length. It estimated the numerically determined wavenumber dependence with a worst-case error of approximately a factor of five. The distribution of the diffraction intensity on a sphere of constant wavenumber was also examined. Finally, the correlation of diffraction intensities in the wavenumber space was assessed. This analysis enabled the estimation of an attainable structural resolution as a function of the incident X-ray intensity density and the volume and length of a target molecule, even in the absence of molecular coordinates.

Key words: X-ray diffraction intensity function, single particle imaging, XFELs, attainable structural resolution, standard model protein

X-ray free-electron lasers (XFELs) generate intense X-ray laser with very short pulses [1,2]. They have ushered in a new era of biological sciences in which the structures and dynamics of crystalline and single-particle samples are elucidated. In nanocrystal diffraction imaging, three-dimensional (3D) structures are built with intense X-ray laser irradiation at a ~10 fs pulse width [3–5]. This configuration realizes "proof before destruction" [6,7]. Single-particle imaging (SPI) is the most challenging but important method for revealing biomolecule structures and dynamics [8]. Typically, a single target molecule of unknown molecular orientation is injected into a vacuum. The X-ray strikes the sample and generates a coherent diffraction pattern that is

Corresponding author: Atsushi Tokuhisa, RIKEN Cluster for Science and Technology Hub, 6-3-5, Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan.
e-mail: tokuhisa@riken.jp



◀ ***Significance*** ▶

X-ray free-electron lasers generate intense X-rays with very short pulses and help elucidate the structures of single-biomolecule samples. Single-particle imaging (SPI) is important for revealing biomolecule structures; however, owing to technical challenges, the SPI resolution is not high. For further progress, it is useful to reasonably know the incident X-ray intensity to achieve the desired resolution. This study clarified how the characteristics of X-ray diffraction are determined from the molecular structure to estimate the attainable resolution even for molecules with unknown coordinates.

observed with a two-dimensional (2D) Charge-Coupled-Device (CCD) detector. To construct the average 3D ensemble structure, multiple measurements must be made using new, identical samples with nearly the same conformational states but in different molecular orientations. A three-step strategy for 3D reconstruction from noisy 2D diffraction patterns has been proposed [9]. In the first step, similar diffraction patterns are grouped and averaged to improve the signal-to-noise (S/N) ratio. In the second step, the 3D diffraction intensity is constructed by aligning the 2D diffraction patterns in the k-space. In the final step, phases not obtained by experimental observation are retrieved by oversampling with the phase-retrieval algorithm [10] adapted to the 3D diffraction intensity function.

Several SPI trials with XFELs have used large living cells or rigid viral samples. Consequently, low-resolution datasets with few diffraction patterns and small scattering angles were obtained such as the 2D projection of a living cell from a diffraction pattern [11] followed by the successful reconstruction of the 3D structure of a virus [12]. This process was accomplished by combining 200 diffraction images, constructing a 3D scattering intensity function in the k-space, and applying the phase-retrieval algorithm.

The structural resolution is low and there are technical challenges. However, theoretical investigation of the attainable structural resolution may help achieve higher resolutions and enhance the applicability and potential of SPI with XFELs. The actual X-ray diffraction intensity of a single biological molecule is extremely weak and susceptible to severe quantum shot-noise. In other words, the attainable structural resolution is determined by X-ray diffraction intensity, and the property of X-ray diffraction intensity depends on the incident X-ray intensity and the target molecule. Hence, it is useful to construct a theory that helps to reasonably and reliably know the necessary incident X-ray intensity to achive the desired resolution, even for molecules whose coordinates are unknown as a function of molecular size.

In an earlier study [13], we reported that the resolutions of spheroid globular molecules are determined from the following characteristics of the diffraction intensity function: $i(\mathbf{k})$ in the wavenumber $\mathbf{k}$ space, (1) average $\bar{i}(k)$ of $i(\mathbf{k})$ on a sphere $|\mathbf{k}|=k$ of radius k, (2) distribution of $i(\mathbf{k})$ on the sphere, and (3) correlation length of $i(\mathbf{k})$ on the sphere. The aim of this research was to clarify how these characteristics of X-ray diffraction intensity are determined from the 3D structure of a molecule. As to (1), it will be shown that the value of the average function $\bar{i}(k)$ is determined up to about a factor of five by specifying molecular 3D structure in terms of only two parameters, volume V and length L of the molecule. As to (2), the distribution function will be shown to be independent of the molecule. As to (3), the correlation length will be shown to depend only on length L. The attainable structural resolution is determined by the characteristics of $i(\mathbf{k})$ that can be specified by V and L. The results of this analysis were used to derive an attainable structural resolution as a function of the incident X-ray intensity density and the volume and length of a target molecule even if there are no atomic coordinates for the molecule.

As a secondary benefit of the characteristics of $i(\mathbf{k})$, the high-throughput analysis can be realized to estimate the molecular volume and length to evaluate the structure and dynamics of protein from the average function $\bar{i}(k)$ that can be approximately obtained by taking the radial average of the noisy two-dimensional diffraction image data. Additionally, the average $\bar{i}(k)$ of $i(\mathbf{k})$ on a sphere $|\mathbf{k}|$=k of radius k is essentially the quantity measured in small-angle solution scattering. Therefore, it is expected that the results for the characteristics of the average $\bar{i}(k)$ obtained herein are applicable for this widely used experimental method as well as for the estimation of molecular shape with a relatively low calculation cost.

Methods

Calculation of electron density of $\rho(x)$ and X-ray diffraction intensity of $i(k)$

The focus was the X-ray diffraction intensity of a single biomolecule or its complex (hereafter, a molecule) with roughly spherical shape. The diffraction intensity of s($\mathbf{k}$), expressed as photons arriving at each pixel on the detector, is given by the following expression:

$$\mathrm{s}(\mathbf{k}) = \mathrm{I_i r_{ce}^2}\omega i(\boldsymbol{k}) \tag{1}$$

where $\mathrm{I_i}$ is the X-ray incident intensity density in [photons/pulse/µm^2], $\mathrm{r_{ce}}$ is the classical electron radius, and $\omega=(\lambda/\sigma L)^2$ is the solid angle, where λ denotes the X-ray wavelength, L is the molecular size, and σ denotes the linear oversampling ratio. In the experiment, ω can be derived from the detector pixel size and the distance between the sample and detector. Molecular 3D structure is described by its electron density $\rho(\boldsymbol{x})$, structure factor $F(\boldsymbol{k})$, and diffraction intensity density $i(\boldsymbol{k})$:

$$F(\boldsymbol{k}) = \int d\boldsymbol{x}\rho(\boldsymbol{x})\,exp\,(-2\pi i\boldsymbol{k}\cdot\boldsymbol{x}) \tag{2}$$

$$\begin{aligned} i(\boldsymbol{k}) &= |F(\boldsymbol{k})|^2 \\ &= \int d\boldsymbol{x}_1 d\boldsymbol{x}_2\,\rho(\boldsymbol{x}_1)\rho(\boldsymbol{x}_2)\,exp\{-2\pi i\boldsymbol{k}\cdot(\boldsymbol{x}_1-\boldsymbol{x}_2)\} \end{aligned} \tag{3}$$

To perform a computational study, it was first necessary to calculate the electron density function $\rho(\boldsymbol{x})$ on a grid point of width 0.1 Å. The atomic coordinates of the molecule were acquired from the Protein Data Bank (PDB) [14]. Gaussian electron density functions were assumed for each atom listed in the *International Tables for Crystallography* [15] using the following equation:

$$\rho(\boldsymbol{x}) = \sum_a c_a\left(\frac{1}{2\pi\sigma_a^2}\right)^{3/2} exp\left(\frac{(\boldsymbol{x}-\boldsymbol{x}_a)^2}{2\sigma_a^2}\right) \tag{4}$$

It was assumed that the total electron density of a protein molecule was the sum of the electron densities of the constructed atoms. The electron density of each atom is approximated from the sum of several isotropic Gaussian distributions centered on the nucleus. The subscript a is a serial number for the Gaussian distribution. One atom is represented by several Gaussian distributions. $F(\boldsymbol{k})$ was also directly calculated as follows:

$$F(\boldsymbol{k}) = \sum_a c_a exp(-2\pi i \boldsymbol{k} \cdot \boldsymbol{x}_a) exp(-2(\pi k \sigma_a)^2) \tag{5}$$

Seven spheroid globular molecules of various molecule sizes were selected from the PDB. The selected codes were 2C9R, 2LZM, 1TV4, 1E7U, 1EPW, 1DP0, and 1KYI. The electron density function ρ(x) and the X-ray diffraction intensity function i(k) were determined for these seven proteins.

Definition of the molecular surface and region of Ω

The average values of the electron density $\langle\rho\rangle$ and squared electron density $\langle\rho^2\rangle$ in the molecular region of Ω play important roles. Because the electron density rapidly decays in the shallow region from the molecular surface, these numerical values closely relate to the definition of molecular surface. Thus, it was necessary to define the molecular surface of the proteins carefully. For proteins with atomic coordinates cited in the PDB, the space wherein the protein atoms existed was divided into the cubic lattice with a lattice constant of 0.1 Å and the electron density ρ was calculated as described above. A proper cutoff for ρ_c was selected and each cube was designated as real if $\rho \geq \rho_c$ or empty if $\rho < \rho_c$. Cubes within 1.4 Å (water molecule radius) of any real cube were considered to be in the hydrated protein interior. All others were regarded as being in the hydrated protein exterior. Cubes within 1.4 Å of any cube in the hydrated protein exterior were considered to be in the protein exterior. All others were regarded as being in the protein interior. The contact surface between the protein interior and protein exterior was defined as the molecular surface. Molecular surface is essentially the same as the Connolly contact surface except for the following points. Generally, the region around 1.4 Å from the van der Waals (VDW) radius of each atom is defined as the protein surface. However, in our method, the criterion is defined by giving an electron density threshold value of ρ_c to define a proper cutoff that is suitable for protein molecule instead of the VDW radius. By adding this improvement of molecular surface for protein, the universal behavior of the electron density product function of the protein can be elucidated.

Results and Discussion

Factors determining wavenumber dependence of the X-ray diffraction intensity

To determine the wavenumber dependence of the diffraction intensity density, we examined an average $\bar{\imath}(k)$ of $i(\mathbf{k})$ on a sphere $|\mathbf{k}|=k$ of radius k. Hereafter, we refer to this quantity as the radial diffraction intensity density. It is obtained by replacing the last factor of the right-hand side of equation (3) by its average on the sphere $k=|\mathbf{k}|$:

$$\bar{\imath}(k) = \int d\boldsymbol{x}_1 d\boldsymbol{x}_2 \rho(\boldsymbol{x}_1)\rho(\boldsymbol{x}_2) sinc(2k|\boldsymbol{x}_1 - \boldsymbol{x}_2|) \tag{6}$$

$$\text{where } sinc\, x = \frac{\sin \pi x}{\pi x} \tag{7}$$

From equation (6), we see that

$$\bar{\imath}(0) = Q^2 \tag{8}$$

where Q is the total number of electrons in the molecule. Equation (6) is transformed as follows:

$$\bar{\imath}(k) = \int dr q(r) sinc(2kr) \tag{9}$$

$$q(r) = \int d\boldsymbol{x}_1 d\boldsymbol{x}_2 \rho(\boldsymbol{x}_1)\rho(\boldsymbol{x}_2) \delta(r - |\boldsymbol{x}_1 - \boldsymbol{x}_2|) \tag{10}$$

Thus, the radial diffraction intensity density is related to equation (10) and may be expressed to a very good approximation as the product of two factors. One relates to the external shape of the target molecule while the other is associated with atom packing inside the molecule. To the end, we must define the molecule surface. A reasonable choice for this purpose is essentially the Connolly molecular surface [16]. The region enclosed by the molecular surface is the molecular region of Ω. The integrations with respect to $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, except for the autocorrelation part ($\boldsymbol{x}_1 = \boldsymbol{x}_2$) in equation (10), lie within the region of Ω. The following pair distribution function relates to the external shape of the target molecule:

$$P(r) = \int_\Omega d\boldsymbol{x}_1 \int_\Omega d\boldsymbol{x}_2 \delta(r - |\boldsymbol{x}_1 - \boldsymbol{x}_2|), \quad \boldsymbol{x}_1 \neq \boldsymbol{x}_2 \tag{11}$$

This equation expresses the number of pairs of points separated by distance r that fit in the molecule and is normalized as follows:

$$\int_0^L dr P(r) = V^2 \tag{12}$$

Where the molecular size of L is defined as the maximum distance between two points in the molecule and V is the volume of the molecular region Ω.

$$R(r) = \frac{q(r)}{P(r)} \tag{13}$$

Equation (13) defines the expected electron density product for a pair of points separated by distance r. This quantity should be more or less the same for most biomacromolecules of similar atomic composition. Thus, it is henceforth referred to as the universal electron density product function.

From equation (13), we have:

$$q(r) = R(r)\,P(r) \tag{14}$$

Therefore, equation (10) is now expressed as a product of the pair distribution function $P(r)$ (related to the external shape of the molecule) and the universal electron density product function $R(r)$ (associated with the atom packing within the molecule).

Pair distribution function of *P*(*r*)

We start with the simplest idealized external molecular shape, namely, a spherical molecule of radius a. A basic calculation yields the following result:

$$P(r) = \frac{V^2}{2a} f(t) \tag{15}$$

$$V = \frac{4\pi a^3}{3} \tag{16}$$

$$t = \frac{r}{2a} \tag{17}$$

$$f(t) = 12t^2(t-1)^2(t+2) \tag{18}$$

where V is the volume of the sphere and t is a dimensionless length between two points within the molecule expressed as a sphere diameter. The range of this variable is between 0 and 1. The function $f(t)$ is related to the shape of the molecule but not its size. The molecule size influences $P(\mathbf{r})$ via the first factor on the right-hand side of equation (15), namely, v^2/2a.

For essentially spherical globular molecules, we assume that equations (15), (16), and (17) hold for the molecular volume V and length a and that these are derived from the volume via equation (16). V is reasonably well estimated from the 3D molecular structure. Thus, we assume that the length a may be determined fairly accurately if there is some indication that the molecule is globular. As we are focusing on biological macromolecules or their complexes, we assume that a ranges from ten to a few hundred angstroms. Here, we designate f(t) as the molecular shape function. It depends on the shape but not the size of the molecule. The variable t ranges from 0 and $\xi = L/2a$ where L is the maximum distance between two points in the molecule. Therefore, ξ describes the deviation of the external molecular shape from an ideal sphere and we refer to it as the (first) shape parameter. The molecular shape function is normalized as follows:

$$\int_0^{\xi} f(t)\,dt = 1 \tag{19}$$

Figure 1 shows values of this function obtained for several molecules listed in the PDB. Here, ξ ranges between 1.5 and 2.2. This rather narrow range quantitatively represent globular molecules with approximately spherical shapes. Figure 1 indicates that there is relatively little variation in the molecular shape function. For this reason, we introduced a standard model protein molecule characterized in terms of the given values of V and $f(t)$:

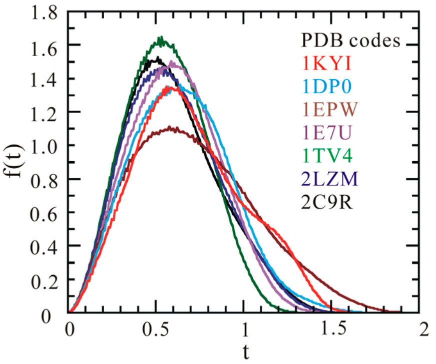


Figure 1 Molecular shape function calculated for seven roughly spherical molecules selected from the PDB. The PDB codes of the molecules are shown.

$$f(t) = \frac{105}{\xi^7} t^2(t-\xi)^4 \tag{20}$$

As before, $\xi = L/2a$ and this function satisfy equation (19). Comparisons of this model with real functions are shown in Figure 2. The fit between the model and the real function was excellent for the T4 lysozyme molecule. In contrast, clear deviations between the two curves were observed for the HslUV complex. In the case of T4 lysozyme, i.e., globular proteins, the real function of f(t) has a single peak. However, in the case of HslUV complex, the real function of f(t) has multiple peaks and shoulders. This behavior in the real function of the HsUV complex is thought to be derived from the topographical features, which are elongated multidomain cylindrical 3D shapes and large hollow interior spaces. However, this discrepancy was, in fact, relatively small considering the peculiar 3D shape of the HslUV complex. We endeavored to establish how well the real function can be replaced by the one-parameter model so that we could evaluate various aspects of diffraction. The model resembles the real functions in the following ways. (1) The increase in the origin is proportional to t^2 as theoretically required according to equation (24) below. (2) The model decreases at a faster rate at the maximum t than that at the origin. (3) The maximum value ~2.30/ξ occurs at $t = \xi/3$. The similarity of the standard model protein molecule to real protein molecules suggests that $\bar{\imath}(k)$ for the latter may be approximated with reasonable accuracy by a function with only two parameters, namely, V and ξ.

Here, we focused on the functional form of $P(r)$ around $r < 2$ Å to derive a simplified expression. In this special region, the electron density product function $R(r)$ is before asymptotic to $\langle\rho\rangle^2$. The value of $\langle\rho\rangle^2$ corresponds to the value of a sufficiently deep inner molecule where the space correlation of electron density vanishes, as demonstrated in the

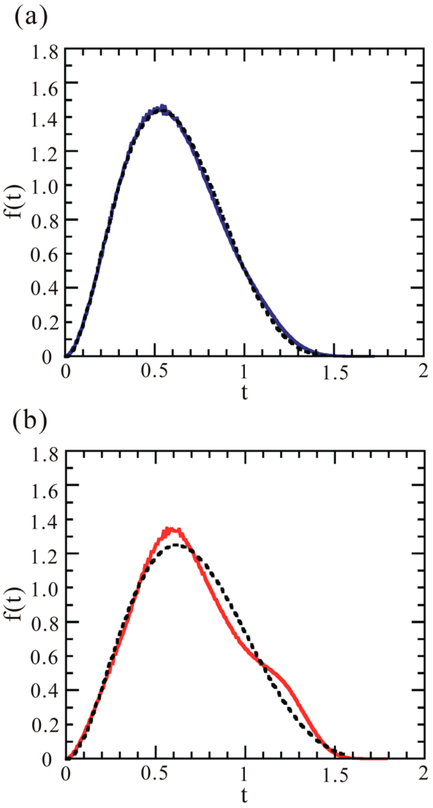


Figure 2 Comparison of the numerically obtained shape function (solid line) with its model molecular shape function (equation (20)) (broken line) in which ξ was adjusted for the best fit. (a) Lysozyme (Weaver & Matthews, 1987) had $\xi_{adjusted}=1.60$; [17] (b) The HslUV complex (Sousa *et al.*, 2000) had $\xi_{adjusted}=1.84$ [18].

following section. This behavior is due to the nature of electron density distribution function near the molecular surface region that directly expresses the electron density distribution near the nucleus of isolated atoms. As shown later, to derive an analytical model form of $\bar{\imath}(k)$ for such a special region, $P(r)$ can be transformed as follows by approximating the surface of the molecule using a plane:

$$\begin{aligned} P(r) &= \int_\Omega d\boldsymbol{x}_1 \int_\Omega d\boldsymbol{x}_2\, \delta(r-|\boldsymbol{x}_1-\boldsymbol{x}_2|) \\ &= \int_\Omega d\boldsymbol{x}_1 \left(\int_{\text{all}} d\boldsymbol{x}_2 - \int_{\bar{\Omega}} d\boldsymbol{x}_2\right) \delta(r-|\boldsymbol{x}_1-\boldsymbol{x}_2|) \\ &= 4\pi r^2 V - \pi r^3 S \end{aligned} \tag{21}$$

where $\bar{\Omega}$ is the volume outside the molecular region, *all*=all space, and S is the surface area of the molecule. We introduced the approximation that for r in this range, the surface is estimated to be flat in the second integration in equation (21):

$$\pi r^3 = \int_\Omega d\boldsymbol{x}_1 \int_{\bar{\Omega}} d\boldsymbol{x}_2\, \delta(r-|\boldsymbol{x}_1-\boldsymbol{x}_2|)$$

$$\text{[where } \boldsymbol{x}_1=(x_1,0,0) \text{ with } x_1<0, \text{ and } \boldsymbol{x}_2=(x_2,y_2,z_2) \text{ with } x_2>0,\ -\infty<y_2<\infty,\ -\infty<z_2<\infty] \tag{22}$$

We introduced the parameter ς to indicate the deviation of the external shape of the molecule from the ideal sphere:

$$S = 4\pi a^2 \varsigma \tag{23}$$

where a is defined by equation (16). We referred to ς as the second shape parameter. V, S, ξ, and ς are calculated for the molecules listed in Figure 1 based on their atomic coordinates in the PDB. The calculations for each molecule are shown in Table 1. The value of ς ranged between 1.5 and 5.0 and is expected to increase with the molecular size, up to 5.0. While ξ describes the deviation of the external molecular shape from an ideal sphere, ς pertains to the smoothness of the molecular surface. Thus, $P(r)$ can be expressed by:

$$\begin{aligned} P(r) &= 4\pi V r^2 \left(1-\frac{3\varsigma r}{4a}\right) \\ &= \frac{64\pi^2}{3} a^5 t^2 \left(1-\frac{3\varsigma t}{2}\right), \quad t=\frac{r}{2a} \end{aligned} \tag{24}$$

The quality of this expression has been numerically validated for several protein molecules (Fig. 3). Unlike ξ, ς has negligible impact on the gross feature of $\bar{\imath}(k)$. By expressing the contribution of $P(r)$ using an approximate analytic form,

Table 1 Calculations of electron density $\rho(x)$ in molecule region Ω for selected seven proteins assuming $\rho_c=0.018$ ((*#of electrons*)/Å^3)

Molecules	CopC	T4lysozyme	MtmB	PI3K	BoNT/B	β-galactosidase	HslUV
PDB ID	2C9R	2LZM	1TV4	1E7U	1EPW	1DP0	1KYI
Molecular volume V [Å^3]	12,160.92	22,629.17	60,544.59	124,989.48	186,496.18	563,562.64	815,665.40
Molecular surface area S [Å^2]	3,867.93	6,090.04	12,206.63	31,005.53	43,263.53	113,394.92	195,298.05
Total electrons in Ω	5,336.34	9,947.23	26,621.19	53,496.38	80,121.25	244,714.74	347,022.83
$\langle\rho\rangle$	0.4388	0.4396	0.4397	0.4280	0.4296	0.4342	0.4254
$\langle\rho^2\rangle$	18.67	18.65	19.39	18.07	17.96	18.48	17.87
$\langle\rho\rangle^2$	0.1926	0.1932	0.1933	0.1832	0.1846	0.1886	0.1810
Molecule size parameter a [Å]	14.27	17.55	24.36	31.02	35.44	51.24	57.96
First shape parameter ξ	1.650	1.718	1.520	1.701	2.178	1.809	1.786
Second shape parameter ς	1.512	1.574	1.637	2.565	2.741	3.437	4.626

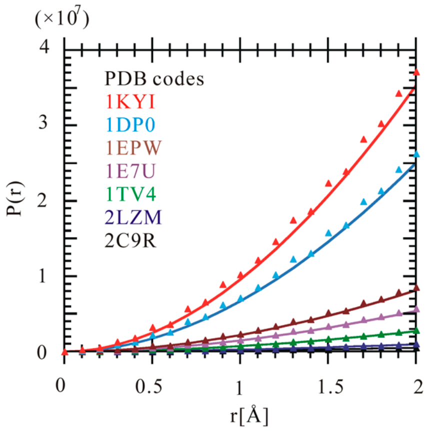


Figure 3 Numerically obtained pair distribution function (triangles) for small values of r fitted with the analytical function of equation (24) (solid line) for the seven molecules listed in Figure 1. The PDB codes of these molecules are shown.

as shown above, we can decompose the original function of $\bar{\imath}(k)$ into two components of $\bar{\imath}_1(k)$ and $\bar{\imath}_2(k)$, with border on r_c=2 Å.

Universal electron density product function of R(r)

$R(\boldsymbol{r})$ is defined by equation (13). Consider an asymptotic form of this factor for $r\rightarrow 0$ and $r\rightarrow\infty$. When distance r becomes larger than a certain value r_c=2 Å, i.e., where the space correlation of electron density vanishes, $R(r)$ rapidly approaches the asymptotic value of $\langle\rho\rangle^2$, with $\langle\rho\rangle=Q/V$, which is the average electron density within the molecular region Ω. Then $r\rightarrow 0$ should be represented by $\langle\rho\rangle^2$. The behavior between 0 and r_c is elucidated by a calculation based on protein atomic coordinate data. The functional form of $R(\boldsymbol{r})$ is expressed as:

$$R(\boldsymbol{r}) = (\langle\rho^2\rangle - \langle\rho\rangle^2)c(r) + \langle\rho\rangle^2 \tag{25}$$

where $c(r)$ is the normalized electron density product function. At r=0 it is unity whereas it vanishes when $r>r_c$. The $\langle\rho^2\rangle$ and $\langle\rho\rangle^2$ and the functional form of $c(r)$ must be determined empirically. Calculations conducted for several proteins consisting only of the standard 20 amino acid residues showed that $R(r)$ of equation (13) is universal and highly accurate, as shown in Figure 4.

As long as $R(r)$ is defined by the right-hand-side of equation (13), our theory is corroborated. $R(r)$, as defined for each molecule by the right-hand-side of equation (13), is approximately a universal function. Based on this understanding, molecular individuality affects the function $q(r)$ which determines $\bar{\imath}(k)$ by equation (9) exclusively through $P(r)$. This remarkable property of $R(r)$ may be significant in terms of the theory of small angle solution scattering. $R(r)$ may be expressed to a certain approximation by a standard model protein molecule characterized by only two parameters. By introducing them and allowing for a certain approximation in $\bar{\imath}(k)$, real protein molecules may be estimated from the standard model protein molecule. Thus, an analytical expression for $\bar{\imath}(k)$ derived for the standard model protein may be used to establish the relationships among the (1) incident X-ray intensity, (2) molecule volume V and length L, and (3) attainable resolution. Furthermore, a standard model protein may help clarify how the individual characteristics of real proteins affect $\bar{\imath}(k)$.

The universal electron density product function $R(r)$ is expected to approach an asymptote $\langle\rho\rangle^2$ when $r>r_c$, where $\langle\rho\rangle$ is the mean electron density within Ω. The mean electron density within the deep protein interior is reasonably well defined. At the molecule surface, however, the electron density sharply drops. The molecular region Ω consists of both the deep interior and surface regions. The surface must be carefully defined so that the mean electron density within Ω coincides with that of the deep interior and, therefore, assumes a universal value. For numerous proteins, the molecular region Ω and the mean electron density $\langle\rho\rangle$ within it are calculated as functions of the selected ρ_c. The $\langle\rho\rangle$ increases with ρ_c and the rate of increase is greater for smaller than larger proteins. It is preferable that $\langle\rho\rangle$ be as universal as possible. It is reasonable to use the value of ρ_c at which the $\langle\rho\rangle$ vs. ρ_c curves for various proteins intersect. Here, we took ρ_c=0.018 (= (*#of electrons*)/Å^3) as the most appropriate value. Even at this ρ_c, $\langle\rho\rangle$ is not strictly universal; rather, $\langle\rho\rangle$=0.434±0.006 ((*#of electrons*)/Å^3). As an independent verification of our selection of ρ_c, we calculated the standard Connolly contact surface based on van der Waals radii and the mean electron density for proteins comprising only the standard 20 amino acid residues and good structural resolution in the PDB. Thence, we obtained $\langle\rho\rangle$=0.477±0.023 ((*#of electrons*)/Å^3), which almost aligns with those determined using the selected ρ_c.

The $\langle\rho^2\rangle$ and $\langle\rho\rangle^2$ determined for several protein molecules were $\langle\rho\rangle^2$=0.188±0.005 and $\langle\rho^2\rangle-\langle\rho\rangle^2$=18.22±0.50 ((*#of electrons*)^2/Å^6). The narrow distribution of their means indicates that they had minimal effect on $\bar{\imath}(k)$. To calculate $\bar{\imath}(k)$ for the standard model protein, we took their means as the universal values. The functional forms of $c(r)$ calculated for several proteins are shown in Figure 5. As the curves are almost indistinguishable from each other, collectively they almost appear as a single line. The universal function $c(r)$ may be approximated from the following expression:

$$c(r) = c\,exp\,(-\eta_1 r) + (1-c)\,exp\,(-\eta_2 r) \tag{26}$$

where c=0.931, η_1=13.4 Å^{-1}, and η_2=2.19 Å^{-1}. From Figure 5, we see that r_c~2 Å and equation (26) confirms this estimation.

The electron density product functions asymptotically decay from the mutually correlated value of ⟨r^2⟩ to the uncorrelated value of ⟨r⟩^{2}. The distance to decay of the uncorrelated value of ⟨r⟩^{2} is 2 Å, mainly contributed from the molecular surface region. The model function of $c(r)$

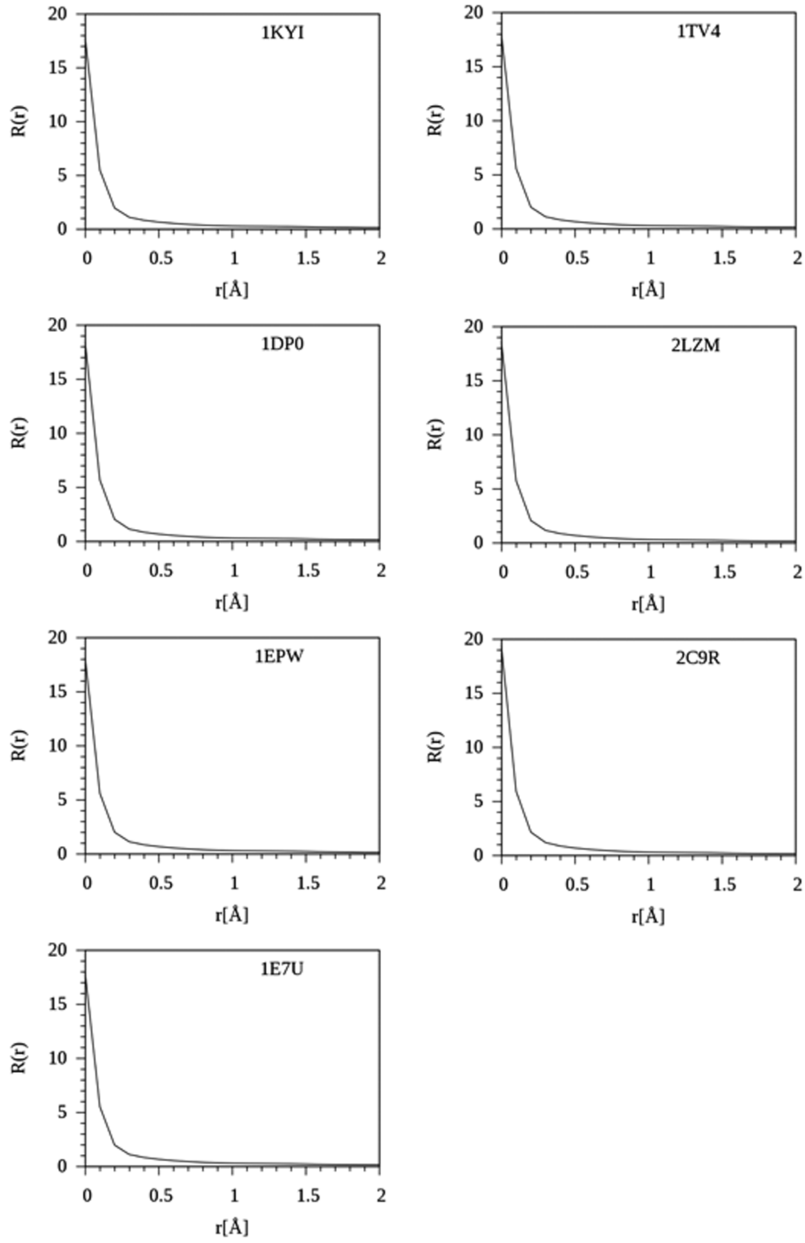


Figure 4 Electron density production function $R(r)$ obtained numerically for the seven molecules listed in Figure 1.

represents the decay function provisionally by superimposing two exponential functions. It can be understood as the contribution of the sharp decay term found in the region between 0 to 0.3 Å in the original function of $R(r)$ and the contribution from the slow decay term found in 2 Å. We estimated three fitting parameters (c, η_1, η_2) for this model from seven proteins. The parameters with dimensions inverse to that of the distance [Å^{-1}] η_1 and η_2 make the 2 Å physical scale of the electron density function for standardization at different scales. The η_1 is a standardization parameter to express the contribution of the sharp decay term using the exponential function. On the other hand, η_2 is a standardization parameter to express the remaining contributions in the region from 0 to 2 Å. The parameter c represents the contribution rate of each exponential function.

The molecular region Ω is calculated by counting the number of surface small cubes (with 0.1 Å sides) in the protein interior with ≥1 neighboring small cubes in the protein exterior. These values are also calculated for the radii of spheres of carbon, nitrogen, and oxygen. It was found that 1.16 multiplied by the number of surface small cubes is the theoretical surface area of the spheres due to the discretization with 0.1 Å. For protein molecules, then, we multiply the same factor to convert the number of surface small cubes into the surface area.

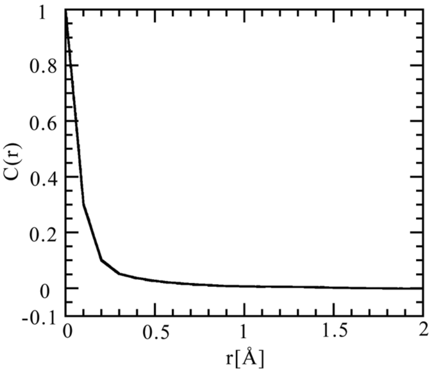


Figure 5 Electron density production function $c(r)$ from equation (26) obtained numerically for the seven molecules listed in Figure 1. As the curves are nearly identical, they collectively resemble a single line.

Radial diffraction intensity density of $\bar{\imath}(k)$

By inserting equation (25) into equations (9) and (14), we obtain:

$$\bar{\imath}(k) = \bar{\imath}_1(k) + \bar{\imath}_2(k) \tag{27}$$

$$\bar{\imath}_1(k) = \langle\rho\rangle^2 \int_0^{\infty} dr\, P(r)\, sinc\,(2kr) \tag{28}$$

$$\bar{\imath}_2(k) = (\langle\rho^2\rangle - \langle\rho\rangle^2) \int_0^{r_c} dr\, c(r)\, P(r)\, sinc\,(2kr) \tag{29}$$

From equation (12), we get:

$$\bar{\imath}_1(0) = \langle\rho^2\rangle V^2 = Q^2 \tag{30}$$

In equation (28), we express $P(r)$ using equations (15), (16), and (17). Thus, we have:

$$\bar{\imath}_1(k) = \langle\rho\rangle^2 V^2 \int_0^{\varsigma} dt\, f(t)\, sinc\,(4kat) \tag{31}$$

In equation (29), we express $P(r)$ using equation (24). Then, we derive:

$$\bar{\imath}_2(k) = 4\pi(\langle\rho^2\rangle - \langle\rho\rangle^2)V \int_0^{r_c} dr\, c(r) r^2 \left(1 - \frac{3\varsigma r}{4a}\right) sinc\,(2kr) \tag{32}$$

In this equation, the shape of a protein molecule is characterized only by ς/a.

The $\bar{\imath}_2(k)$ at $k=0$ does not vanish but is much smaller than $\langle\rho^2\rangle V^2 = Q^2$. Therefore, $\bar{\imath}(0)$ is calculated using equations (27) and (28) (that is, equation (30)) and equation (29) does not satisfy equation (8). This deviation is a consequence of identifying the mean electron density within the deep protein deep interior with $\langle\rho\rangle$, namely, the mean electron density within Ω. As the electron density is relatively lower on the molecule surface, the former should be slightly larger than $\langle\rho\rangle$. This difference accounts for the aforementioned deviation but was neglected as the actual deviation was small.

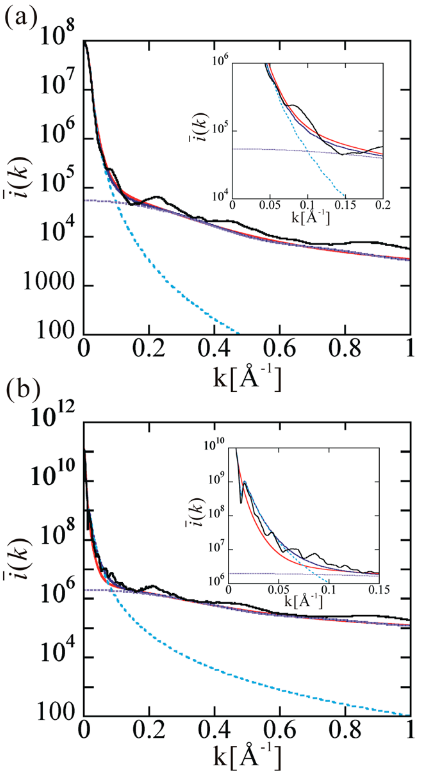


Figure 6 Diffraction intensity density $\bar{\imath}(k)$ obtained by three different methods. Line 1 (black solid line): $\bar{\imath}(k)$ obtained numerically from the atomic coordinates. Line 2 (blue solid line): $\bar{\imath}(k)$ obtained from equations (27), (31), and (32); $\bar{\imath}_1(k)$=(blue dotted line), $\bar{\imath}_2(k)$=(purple dotted line). Line 3 (red solid line): $\bar{\imath}(k)$ approximated by the analytical form of equation (37).

Figure 6 shows $\bar{\imath}(k)$ for lysozyme and HslUV complex obtained by direct calculation using their atomic coordinates. The $\bar{\imath}(k)$, $\bar{\imath}_1(k)$, and $\bar{\imath}_2(k)$ were derived from equations (27), (31), and (32), respectively, and $f(t)$ (Fig. 1) and $c(r)$ (Fig. 3) are used. The $\bar{\imath}_1(k)$ is calculated from equation (31) for an analytical polynomial function fitted to $f(t)$. Figure 6 shows an approximate analytical form of $\bar{\imath}(k)$.

The minor difference between the two $\bar{\imath}(k)$ arose from the values of the universal $\langle\rho^2\rangle$ and $\langle\rho\rangle^2$ in equations (31) and (32) and the universal function $c(r)$ in equation (32). The most conspicuous difference is the "waving" behavior of $\bar{\imath}(k)$ obtained by direct calculation. The first "waving" behavior for the HslUV complex at $k \cong 0.013$ (Å^{-1}) was reproduced in $\bar{\imath}(k)$ of equation (27). It may have been a consequence of the shoulder of $f(t)$ at $\sim t=1.0$. The other "waving" behaviors at the larger k were not reproduced in $\bar{\imath}(k)$ of equation (27) because the behavior of the $c(r)$ function is slightly different between the model function and real function in the long-distance of $r>2$ Å. In the model function, it is monotonously decreasing, but in the real function, it has a slightly waving behavior. In our preliminary work, the waving behavior was reproduced when the behavior of the $c(r)$ function was replaced up to $r=2$ Å and $r=7$ Å. However, the accuracy obtained by truncating the $c(r)$ function at $r=2$ Å suffices for the purposes of this analysis.

Figure 6 shows that the contributions of $\bar{i}_1(k)$ and $\bar{i}_2(k)$ to $\bar{i}(k)$ become negligible at wide and narrow k ranges, respectively. Here, we used $P(r)$ in the range of $0<r<r_c$ in equations (31) and (32). This leeway was permissible because of the observed behaviors of $\bar{i}_1(k)$ and $\bar{i}_2(k)$. The analytically fitted $f(t)$ function may differ from equation (24) in a small range of t and, by extension, r, as the fitting is usually performed to obtain an overall uniform overall best fit in the full range of t. The behavior of $f(t)$ in a narrow t range affects the behavior of $\bar{i}_1(k)$ in the wide k range. As long as the analytically fitted $f(t)$ retains reasonable function in the narrow t range, it will be reflected by negligibly small $\bar{i}_1(k)$ in the wide k range. Therefore, we can use $f(t)$, which is slightly sloppy at the lower end of the t range.

Here, we clarify the limit of structural resolution due to quantum noise as a function of the incident X-ray intensity and target molecule size and shape. To this end, we introduced the standard model protein defined by equation (20). By substituting equations (15) and (20) into equation (28), we have:

$$\bar{i}_1(k)=\langle\rho\rangle^2V^2\frac{840(15\sin u-3u\cos u-12u+u^3)}{u^7},\quad u=4\pi\xi ka \tag{33}$$

As k (and, by extension, u) approach zero, the right-hand-side of this equation approaches $\langle\rho\rangle^2V^2$ as required by equation (30). The ka in equation (33) is the wavenumber scaled by the radius of an equivalent sphere a. When k is the wavenumber of the structural resolution k_R, k_Ra is the number of independent descriptive structural elements along the radius of the equivalent sphere a. In order for the single molecule imaging method to be informative, this number must be ≥10 and preferably 100. Therefore, the upper bound of u in equation (33) must be 150–1,500. For large u, equation (33) is approximated by:

$$\bar{i}_1(k)=\langle\rho\rangle^2V^2\frac{840}{u^4},\quad u=4\pi\xi ka \tag{34}$$

Thus, $\bar{i}_1(k)$ decreases with increasing u.

We consider the functional form of $\bar{i}_2(k)$ when equation (26) is substituted into equation (32). The integration gives the following:

$$\bar{i}_2(k)=3(\langle\rho^2\rangle-\langle\rho\rangle^2)V^2\left[\begin{array}{l}c\left(2\eta_1a+\frac{3\varsigma}{2}\right)\frac{1}{((\eta_1a)^2+(2\pi ka)^2)^2}-\frac{6\varsigma c(\eta_1a)^2}{((\eta_1a)^2+(2\pi ka)^2)^3}\\+(1-c)\left(2\eta_2a+\frac{3\varsigma}{2}\right)\frac{1}{((\eta_2a)^2+(2\pi ka)^2)^2}-\frac{6\varsigma(1-c)(\eta_2a)^2}{((\eta_2a)^2+(2\pi ka)^2)^3}\end{array}\right] \tag{35}$$

In view of the aforementioned ranges for the variables and parameters in this equation, this expression is approximated with an error of a few percentage points as follows:

$$\bar{i}_2(k)=3(\langle\rho^2\rangle-\langle\rho\rangle^2)V^2\left[\frac{2c\eta_1a}{((\eta_1a)^2+(2\pi ka)^2)^2}+\frac{2(1-c)\eta_2a}{((\eta_2a)^2+(2\pi ka)^2)^2}\right] \tag{36}$$

The contribution of the term proportional to the surface area S in equation (21) to $\bar{i}_2(k)$ can be neglected if an error of a few percentage points is acceptable. Summarizing the above, we obtain:

$$\frac{\bar{i}(k)}{Q^2}=\frac{840(15\sin u-3u\cos u-12u+u^3)}{u^7}+3\left(\frac{\langle\rho^2\rangle}{\langle\rho\rangle^2}-1\right)\left[\frac{2c\eta_1a}{((\eta_1a)^2+v^2)^2}+\frac{2(1-c)\eta_2a}{((\eta_2a)^2+v^2)^2}\right]$$

$$u=2\xi v,\ v=2\pi ka \tag{37}$$

Figure 6 shows $\bar{i}(k)$ derived from this analysis.

Here, to evaluate the functional form of $\bar{i}(k)$, the evaluation function is defined using the following equation as a function of k between the analytical model function and numerical original function.

$$\mathrm{diff(k)}=\bar{i}_{model}(k)/\bar{i}_{original}(k) \tag{38}$$

The statistics of the evaluation function are presented in Table 2 for case of lysozyme and HslUV complex. In the case of lysozyme, the average value of the evaluation function for k is $\langle\mathrm{diff(k)}\rangle_k=0.789\pm0.199$. In the case of HslUV complex, $\langle\mathrm{diff}(k)\rangle_k=0.753\pm0.269$. It is evident that the value of $\bar{i}(k)$ obtained from the model function is approximately 25% smaller on average than the original function. In the case of T4 lysozyme, the largest difference is that the evaluation function is 0.53 when k is 0.899. In the case of HslUV complex, the worst case is that the evaluation function is 6.46 when k is 0.012. Although different in the case of HslUV complex than in the case of lysozyme, they are derived from the fine structure of waving behavior found in the small angle region of the i (k) function in the case of the HslUV complex. When the distribution of the evaluation function of the HslUV complex was examined, a total of 38

Table 2 Statistics of the evaluation function of diff(k)

	T4 lysozyme	HslUV complex
Maximum of diff(k)	1.548	6.46
k [Å] at the maximum	0.046	0.012
Minimum of diff(k)	0.532	0.346
k [Å] at the minimum	0.899	0.044
Average of diff(k) for k	0.789	0.753
Variance of diff(k) for k	0.0396	0.0724
Standard deviation of diff(k) for k	0.199	0.269

points of k at which the difference is doubled or less than half were found, and all were in the small-angle region where k was less than 0.1. This indicates that the global shape is deviated from the spherical model from the rough approximation of spherical model in the case of HslUV complex.

From the above analysis, the analytical form approximates the calculated $\bar{i}(k)$ within the worst-case errors of factors <2 for lysozyme and ~5 for the HslUV complex. In this expression, only the molecule size a and molecule shape ξ can be considered as parameters. Nevertheless, the equation reproduces the calculated "exact" $\bar{i}(k)$ within a worst-case error of a factor of ~5. Thus, for the purpose of discussing the incident X-ray intensity as a function of molecular size and shape and the attainable resolution, this analytical expression is reliable.

These results indicated that molecular 3D shape may be relatively simply calculated and generated in an analytical form such that the X-ray diffraction intensity is accurately approximated using as few as two parameters i.e. molecular length and volume. The structures and dynamics of even roughly spherical single-particle protein molecules may be reliably estimated based on their external morphology and atomic packing.

Distribution of $i(k)$ on a sphere of radius k

The purpose of this and the next subsections is to numerically clarify the shape of the distribution functions in protein that is important to derive the theoretical equations for the variance value of the diffraction intensity function of σ_c^2. This distribution was calculated for several k of the proteins lysozyme and HslUV complex. For this purpose, ~1.5×10^5 points were randomly sampled with uniform probability on each sphere and $i(\boldsymbol{k})$ calculated at each point was normalized with its mean value. The distributions of these normalized values on each sphere are shown for the HslUV complex in Figure 7. Except at very low k, the distribution was exponential. This subsection succeeded in showing that the distribution on the sphere of $i(\boldsymbol{k})$ is exponential distribution. Wilson (1949) reported that the irregular 3D structures of biopolymers at the atomic level are the origin of this distribution [19].

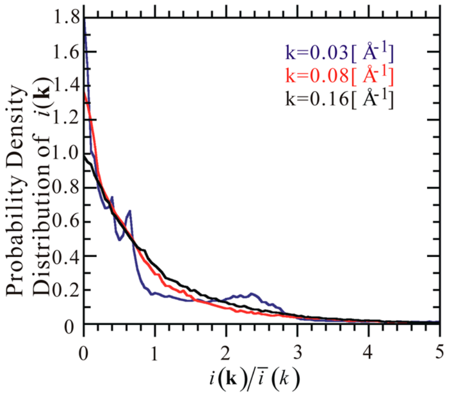


Figure 7 Probability density distribution of the $i(\boldsymbol{k})$ on a sphere $|\mathbf{k}|=k$ of radius k for the HslUV complex. For $k\geq0.16$ Å^{-1}, the distribution was exponential. Data up to and including 0.16 Å^{-1} are shown. Deviation from the exponential distribution increased with decreasing k.

Correlation of $i(k)$ on a sphere $|\mathbf{k}|=k$ of radius k

For the purpose of this analysis, we used the same set of data on each sphere as in the previous section. We calculated $\langle i(\mathbf{k}_1)i(\mathbf{k}_2)\rangle_\delta$ and the average was taken over all pairs of **k** vectors. The angle between them was δ. As shown in Figure 8, except at very low k, the distribution was Gaussian:

$$c_N(\delta)\equiv\frac{\langle i(\mathbf{k}_1)i(\mathbf{k}_2)\rangle_\delta-\langle i\rangle^2}{\langle i^2\rangle-\langle i\rangle^2}=exp\left(-\left(\frac{k\delta}{k_c}\right)^2\right) \tag{39}$$

The correlation length k_c was approximated by:

$$k_c=\frac{1}{L} \tag{40}$$

Thus, the correlation length was associated with the length of the molecule under investigation and was independent of k.

An attainable resolution as a function of the incident X-ray intensity and the size of a target molecule

For further development of SPI using XFEL, it is useful to construct a theory, which helps to reasonably and reliably know the necessary incident X-ray intensity density needed to achieve the desired resolution, even for molecules whose coordinates are unknown. If the functional form of $\bar{i}(k)$ can be predicted with a reasonable reliability only from the minimum amount of information about the molecule, it is possible to estimate an attainable resolution by the described method in an earlier study [13].

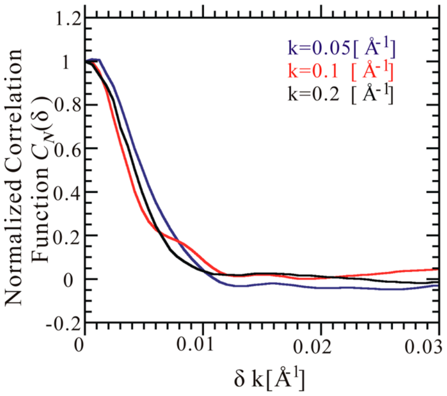


Figure 8 Normalized correlation function $c_N(\delta)$ of equation (39) for the space correlation of $i(\boldsymbol{k})$ on a sphere $|\mathbf{k}|=k$ of radius k for the HslUV complex. δ in the abscissa is a product with k. For $k\geq0.2$ Å^{-1}, the distribution was Gaussian. Data up to and including 0.2 Å^{-1} are shown. Deviations from the Gaussian function slightly increased with decreasing k.

In the above sections, it was shown that an analytic model function $\bar{\imath}(k)$ using a standard model protein, which captures the characteristics of real globular proteins in terms of only two parameters, a and ξ, or equivalently in terms of volume V and length L, can reproduce the behaviour of the real function with the accuracy of a factor of approximately 5. By using this approximate but analytic expression, an attainable structural resolution for the three-step strategy of 3D reconstruction from many noisy 2D diffraction patterns was estimated for a 'molecule', which is characterized by the radius of equivalent sphere a.

Here, the outline of an estimating attainable structural resolution is described briefly as following. Please refer to the article for details. The attainable structural resolution for SPI depends on the 3D structure reconstruction strategy. As it is under development, there are various type of algorithm for 3D reconstruction currently. Here, as an example, the attainable resolution is estimated for the conventional three-step algorithm, i.e., the classifying step, the assembling step, and the phase retrieval step, in brief. In the three-step algorithm, the classification accuracy determines the attainable structural resolution. In the classifying step, the similarity between an arbitrary pair of diffraction images is found by using a correlation function, and the signal to noise (S/N) ratio of the diffraction image is improved by averaging the diffraction images into similar group. In the wide-angle region in the diffraction pattern, the signal is buried in the quantum-shot noise effect because the diffraction intensity is significantly reduced in this region. When a pair of diffraction images are similar, the correlation line appears in the correlation pattern, demonstrating the correlation function of a pair of diffraction images. However, the correlation line is valid in the quantum shot-noise effect in the wide-angle region and cannot be recognized. The wave number at which the quantum shot-noise effect becomes noticeable is denoted as k_N. When the diffraction patterns are averaged within a similar group, the S/N ratio of the averaged diffraction image is improved in the wavenumber region at an angle lower than k_N. In contrast, in the wide-angle wavenumber region outside of k_N, the signal is lost because it is no longer guaranteed that the speckle patterns are similar in the group. Considering the above reasons, the limit wave number of k_N, whose S/N ratio is improved in the diffraction averaging process, is defined as the resolution wave number of k_R. The limit wave number of k_N at which the noise becomes noticeable can be estimated from the degree of noise in the diffraction pattern. The standard deviation of σ_c of the diffraction intensity is used as an index of the degree of noise. In a conservative way, the quantum-shot noise becomes noticeable at a wave number of $\sigma_c = \exp(-1/2) \cong 0.6$ and the correlation line disappears in the correlation pattern. Thus, the attainable resolution can be estimated by the expected mean value and the variance value (or the standard deviation value) of the number of photons observed at the effective pixel. The variance value of the diffraction intensity can be theoretically given by the following equation as a function of $\bar{\mathrm{s}}(\mathrm{k}) = I_i r_{ce}^2 \omega \bar{\imath}(k)$ with equation (1), which is the average value of the diffraction intensity.

$$\sigma_c^2 = \frac{g(\bar{\mathrm{s}}(\mathrm{k}))}{N_\xi} \tag{41}$$

Here,

$$g(x) = \frac{5x^2 + 6x + 1}{x^2} \tag{42}$$

$$N_\xi = 2\pi k L \left[1 - \left(\frac{k\lambda}{2}\right)^2\right]^{1/2} \tag{43}$$

From the above equations, the diffraction intensity can be expressed using the following equation as an inverse function of g (x).

$$\bar{\mathrm{s}}(\mathrm{k}) = g^{-1}(N_\xi \sigma_c^2) \tag{44}$$

From equation (1) and equation (44), the formula for estimating the attainable resolution can be obtained.

$$I_i = \frac{g^{-1}(N_\xi \sigma_c^2)}{r_{ce}^2 \omega \bar{\imath}(k)} \tag{45}$$

By constructing the theory of $\bar{\imath}(k)$ and proposing a "standard protein model" in this study, it can be applied to even a molecule of unknown structure as a parameter of molecular size of a.

The necessary incident X-ray intensity density estimation program constructed here can flexibly cope with the experimental conditions by changing the incident X-ray wavelength (or energy), oversampling ratio, molecular size, and the first shape parameter. As the first example, parameters giving a standard protein model were set as follows: $c=0.931$, $\eta_1=13.4$, $\eta_2=2.19$, $\langle\rho\rangle^2=0.188$, and $\langle\rho^2\rangle=18.41$. In addition, assuming that the size of the effective pixel is Shannon pixel as a standard estimation according to our previous paper, $\omega = \left(\frac{\lambda}{\sigma L}\right)^2$, $\lambda = 1$ Å, $L = 2a\xi$, $\xi = 1.7$, and $\sigma = 1$. If it is needed to estimate under other experimental conditions, it can be responded flexibly by changing these hyperparameters. In addition, the degree of freedom of the linear oversampling ratio can be easily estimated by multiplying the constant factors of σ^2 to the result of the standard estimation shown here. Figure 9 shows the estimated incident X-ray intensity density as a function of molecular size and resolution by numerically solving the g^{-1} using the Newton-Raphson method. The dependence of the result on the value of the parameter ξ is not large. The result in Figure 9 is a good estimation by the verification, as shown in the next section, and therefore, can be used in designing experimental parameters of instruments and targets. The results indicated that if sufficient experimental data can be obtained, high resolution can be achieved by X-ray SPI.

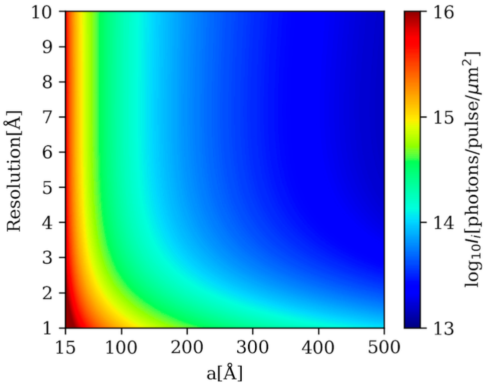


Figure 9 Incident x-ray intensity density [photons/pulse/μm^2] to be used to attain a given structural resolution for a 'molecule' which is characterized by the radius of equivalent sphere *a*.

Verification of estimation accuracy

We verified the estimation accuracy using a standard model protein. In our previous paper [13], the attainable resolution is listed using the numerical real function of $\bar{\imath}(k)$ instead of the model function of $\bar{\imath}(k)$ using the standard model protein for Lysozyme and HslUV complex. First, the case of Lysozyme with L=60 Å is examined. When the incident X-ray intensity density was 1.0×10^{16} [photons/pulse/μm^2], the attainable structural resolution was 1.01 Å, and when it was 5×10^{15} [photons/pulse/μm^2], the attainable structural resolution was 2.08 Å in the previous work. When a standard model protein is used, such as ξ=1.7, the lysozyme molecules correspond approximately to a=18 Å. Under this assumption, the incident X-ray intensity density required to achieve the resolution of 1.01 Å is 2.2×10^{16} [photons/pulse/μm^2], and that required to achieve the resolution of 2.08 Å was estimated to be 8.4×10^{15} [photons/pulse/μm^2] from Figure 9.

The case of HslUV complex with L=200 Å was also estimated. When the X-ray intensity density is 1.0×10^{15} [photons/pulse/μm^2], the attainable structural resolution is 1.82 Å, and when it is 5×10^{14} [photons/pulse/μm^2], the attainable structural resolution is 3.57 Å. When a standard model protein is used, such as ξ=1.7, the HslUV complex corresponds approximately to a=58 Å. Under this assumption, the incident X-ray intensity density required to achieve the resolution of 1.82 Å is 1.4×10^{15} [photons/pulse/μm^2] and that required to achieve the resolution of 3.57 Å was estimated to be 6.1×10^{14} [photons/pulse/μm^2] from Figure 9. In each case, we have succeeded in making estimations for Lysozyme and HslUV complex with high accuracy.

In addition, we verified the estimation accuracy using the actual experimental data.

Here, we verified the estimation accuracy of the constructed model under two SPI experimental conditions. The first was Omono River virus [20] and the experimental conditions described in the paper were as follows: the molecular size is 43.2 nm, the incident X-ray energy is 5.5 kev, and the incident X-ray intensity density is 8.65×10^{11} [photons/pulse/μm^2]. Under the conditions, the resolution estimated from the experimental data was 42.5 Å. Our estimated incident X-ray intensity density required to achieve 42.5 Å resolution using the model function of $\bar{\imath}(k)$ built with λ=2.25 Å, a=216 Å, and ξ=1.0 was 1.3×10^{11} [photons/pulse/μm^2].

The second example was the rice dwarf virus [21] and the experimental conditions described in the paper were as follows: the molecular size is 70.8 nm, the incident X-ray intensity density was not described, but the resolution estimated from the experimental data was 5.9 Å, λ=1.77 Å, and a=354 Å. Using the model with ξ=1.0, i.e., the spherical model, the estimate of the incident X-ray intensity density required to achieve 5.9 Å resolution was 3.9×10^{12} [photons/pulse/μm^2].

In the case of Omono River virus, we succeeded in making estimations with 1/6 times error. In the case of rice dwarf virus, the incident X-ray intensity is not estimated but a 0.1 μm KB mirror is used in experiment. It is inferred that the maximum value of the incident X-ray intensity observed in the experiment of the Omono River virus is not largely different. The maximum value under this condition is 1.9×10^{12} [photons/pulse/μm^2] and if it were this incident X-ray intensity, our model would be estimated about twice as high.

There is a concern that our model does not reproduce the characteristic curves observed in the function of $\bar{\imath}(k)$ derived from the spherical shape. However, we have generally succeeded in providing a good model because we can estimate the necessary incident X-ray intensity density in the correct order even for spherical objects such as viruses.

A one-dimensional radial diffraction intensity function, which can be approximately considered as an $\bar{s}(k)$ function, can be easily calculated using a noisy experimental two-dimensional diffraction data by calculating the radial average. From the function, it is expected that one can estimate the molecular size parameter using the analytical model of $\bar{\imath}(k)$ described in equation (37) constructed herein. In addition, once the one-dimensional radial diffraction intensity function from experiment and molecular size is obtained, the attainable spatial resolution can be directly estimated. Our theory provides a method to estimate the attainable structural resolution from the radial diffraction intensity of $\bar{s}(k)$ and molecular size via equation (41).

Conclusions/Summary

1. The average ($\bar{\imath}(k)$) of the diffraction intensity density function $i(\boldsymbol{k})$ on a sphere $|\mathbf{k}|=k$ of radius k is given by equation (9). The value $q(r)$ is given as shown in equation (14). It is the product of the pair distribution function $P(r)$ defined by equation (11) and the electron density product function $R(r)$ defined by equation (13). $P(r)$ is related only to the external shape of the molecule and is the expected electron density product for a given pair of

points separated by distance r. The latter is related to the atom packing within a molecule.

2. Based on the pair distribution function $P(r)$, the dimensionless molecular shape function $f(t)$ is defined by equations (15), (16), and (17). V is the volume of the molecule. The function $f(t)$ calculated for typical proteins in the PDB had a narrow range. Therefore, it is reasonably well approximated by the one-parameter function of equation (20). A standard model protein is one whose molecular shape function is defined by this function. For $r<2$Å, $P(r)$ is defined by equation (24).
3. The electron density product function $R(\boldsymbol{r})$ is defined by equation (25). The calculated functional form of the normalized electron density product function $c(r)$ is universally applicable to numerous molecules (Fig. 5). This function is defined by equation (26).
4. The approximation applied in the universal electron density product function of equation (25) effectively replicates the calculated $\bar{i}(k)$ (Fig. 6).
5. The analytical expression of $\bar{i}(k)$ in equation (37) is derived for a standard model protein characterized by molecular volume V and length L or by a and ξ, defined by $V=4\pi a^3/3$ and $L=2\xi a$, respectively. This analytical form approximates the calculated $\bar{i}(k)$ with a worst-case error of approximately a factor of five. Thus, for the purpose of discussing the incident X-ray intensity as a function of molecular size and shape and attainable resolution, it can be used reliably.
6. The distribution of $i(\mathbf{k})$ on a sphere $|\mathbf{k}|=k$ of radius k around its mean $\bar{i}(k)$ was exponential.
7. Correlation of $i(\mathbf{k})$ on a sphere $|\mathbf{k}|=k$ of radius k had a Gaussian distribution (equation (39)) and a correlation length defined by equation (40).
8. An attainable resolution as a function of the incident X-ray intensity density and the size of a target globular molecule whose coordinates are unknown using analytical expression of $\bar{i}(k)$ in equation (37).
9. It is established that molecular 3D shape may be relatively simply calculated and generated in an analytical form such that the X-ray diffraction intensity is accurately approximated using as few as two parameters i.e. radius of a sphere with molecular volume, a and ξ describes the deviation of the external molecular shape from an ideal sphere. The structures and dynamics of even roughly spherical single-particle protein molecules may be reliably estimated based on their external shape and atomic packing. This methodology may be applied to the establishment of structure-function relationships for novel biomolecular species, even if they have irregular shapes that are away from ensemble averaged 3D structure in the crystal and even if noisy experimental 2D diffraction patterns of single particles are used.

Acknowledgements

I thank Prof. Nobuhiro Go for his useful comments, advice, assisted in theoretical calculation and research guidance. I also thank Editage (www.editage.jp) for English language editing. This work was supported in part by grants from JSPS KAKENHI (No. 26870852), the Research Complex Promotion Program of the Japan Science and Technology Agency, the FOCUS Establishing Supercomputing Center of Excellence, and MEXT as "Priority Issue 1 on Post-K computer" (Building Innovative Drug Discovery Infrastructure through Functional Control of Biomolecular Systems). This study used the resources of the K computer and the Camphor2 of CRAY XC40 provided by Kyoto University through the High Performance Computing Infrastructure System Research Project (Project I.D. No. hp140121, hp180123, and hp190105).

Conflicts of Interest

The authors declare no conflicts of interest associated with this manuscript.

Ethics Standard

Not applicable

Informed Consent

Not applicable

Author Contribution

A. T. contributed to all computational analysis and interpretation of data, and wrote the manuscript.

References

[1] Emma, P., Akre, R., Arthur, J., Bionta, R., Bostedt, C., Bozek, J., *et al.* First lasing and operation of an ångstrom-wavelength free-electron laser. *Nat. Photonics* **4**, 641–647 (2010).

[2] Ishikawa, T., Aoyagi, H., Asaka, T., Asano, Y., Azumi, N., Bizen, T., *et al.* A compact X-ray free-electron laser emitting in the sub-ångström region. *Nat. Photonics* **6**, 540–544 (2012).

[3] Barends, T. R. M., Foucar, L., Botha, S., Doak, R. B., Shoeman, R. L., Nass, K., *et al. De novo* protein crystal structure determination from X-ray free-electron laser data. *Nature* **505**, 244–247 (2014).

[4] Chapman, H. N., Fromme, P., Barty, A., White, T. A., Kirian, R. A., Aquila, A., *et al.* Femtosecond X-ray protein nanocrystallography. *Nature* **470**, 73–77 (2011).

[5] Suga, M., Akita, F., Sugahara, M., Kubo, M., Nakajima, Y., Nakane, T., *et al.* Light-induced structural changes and the site of O=O bond formation in PSII caught by XFEL. *Nature* **543**, 131–135 (2017).

[6] Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature* **406**, 752–757 (2000).

[7] Solem, J. C. Imaging biological specimens with high-intensity soft x rays. *J. Opt. Soc. Am. B.* **3**, 1551 (1986).

[8] Spence, J. C. H. XFELs for structure and dynamics in biology. *IUCrJ* **4**, 322–339 (2017).

[9] Huldt, G., Szoke, A. & Hajdu, J. Diffraction imaging of single particles and biomolecules. *J. Struct. Biol.* **144**, 219–227 (2003).

[10] Fienup, J. R. Phase retrieval algorithms: a comparison. *Appl. Opt.* **21**, 2758–2769 (1982).

[11] Kimura, T., Joti, Y., Shibuya, A., Song, C., Kim, S., Tono, K., *et al.* Imaging live cell in micro-liquid enclosure by X-ray laser diffraction. *Nat. Commun.* **5**, 3052 (2014).

[12] Ekeberg, T., Svenda, M., Abergel, C., Maia, F. R. N. C., Seltzer, V., Claverie, J.-M., *et al.* Three-dimensional reconstruction of the giant mimivirus particle with an X-ray free-electron laser. *Phys. Rev. Lett.* **114**, 098102 (2015).

[13] Tokuhisa, A., Taka, J., Kono, H. & Go, N. Classifying and assembling two-dimensional X-ray laser diffraction patterns of a single particle to reconstruct the three-dimensional diffraction intensity function: Resolution limit due to the quantum noise. *Acta Crystallogr. A* **68**, 366–381 (2012).

[14] Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., *et al.* The protein data bank. *Acta Crystallogr. D Biol. Crystallog.* **58**, 899–907 (2002).

[15] International tables for crystallography. in *Mathematical, Physical and Chemical Tables.* (Prince, E. ed.) (Kluwer Academic Publishers, Dordrecht, 2004).

[16] Connolly, M. L. Analytical molecular surface calculation. *J. Appl. Crystallog.* **16**, 548–558 (1983).

[17] Weaver, L. H. & Matthews, B. W. Structure of bacteriophage T4 lysozyme refined at 1.7 A resolution. *J. Mol. Biol.* **193**, 189–199 (1987).

[18] Sousa, M. C., Kessler, B. M., Overkleeft, H. S. & McKay, D. B. Crystal structure of HslUV complexed with a vinyl sulfone inhibitor: corroboration of a proposed mechanism of allosteric activation of HslV by HslU. *J. Mol. Biol.* **318**, 779–785 (2002).

[19] Wilson, A. J. C. The probability distribution of X-ray intensities. *Acta Crystallog.* **2**, 318–321 (1949).

[20] Daurer, B. J., Okamoto, K., Bielecki, J., Maia, F. R., Mühlig, K., Seibert, M. M., *et al.* Experimental strategies for imaging bioparticles with femtosecond hard X-ray pulses. *IUCrJ* **4(Pt 3)**, 251–262 (2017).

[21] Munke, A., Andreasson, J., Aquila, A., Awel, S., Ayyer, K., Barty, A., *et al.* Coherent diffraction of single Rice Dwarf virus particles using hard X-rays at the Linac Coherent Light Source. *Sci. Data* **3**, 160064 (2016).