In-plane rotation classification for coherent X-ray imaging of single biomolecules

Kaiqin Chu,[1,*] **James Evans,**[2] **Nina Rohringer,**[3,4] **Stefan Hau-Riege,**[4] **Alexander Graf,**[4] **Matthias Frank,**[4] **Zachary J. Smith,**[1] **and Stephen Lane**[1,4]

[1]*Center for Biophotonics Science and Technology, Sacramento, California 95817, USA*
[2]*Molecular and Cellular Biology, University of California at Davis, Davis, California 95616, USA*
[3]*Max Planck Advanced Study Group, Center for Free-Electron Laser Science, c/o DESY, 22607 Hamburg, Germany*
[4]*Lawrence Livermore National Laboratory, Livermore, California 94550, USA*
**kqchu@ucdavis.edu*

Abstract: We report a new classification scheme with computation complexity well within the capacity of a PC for coherent X-ray imaging of single biomolecules. In contrast to current methods, which are based on data from large scattering angles, we propose to classify the orientations of the biomolecule using data from small angle scattering, where the signals are relatively strong. Further we integrate data to form radial and azimuthal distributions of the scattering pattern to reduce the variance caused by the shot noise. Classification based on these two distributions are shown to successfully recognize not only the patterns from molecules of the same orientation but also those that differ by an in-plane rotation.

OCIS codes: (290.5840) Scattering, molecules; (290.2558) Forward scattering; (110.3200) Inverse scattering; (260.1960) Diffraction theory.

References and links

1. D. Sayre, H. N. Chapman, and J. Miao, "On the extendibility of x-ray crystallography to noncrystals," Acta Crystallogr., Sect. A: Found. Crystallogr. **54**, 323–239 (1998).
2. J. Miao, Charalambous, J. Kirz, and D. Sayre, "Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens," Nature **400**, 342–344 (1999).
3. R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, and J. Hajdu, "Potential for biomolecular imaging with femtosecond x-ray pulses," Nature **406**, 753–757 (2000).
4. S. Marchesini, H. Chapman, S. P. Hau-Riege, R. A. London, A. Szoke, H. He, M. R. Howells, H. Padmore, R. Rosen, J. C. H. Spence, and U. Weierstall, "Coherent x-ray diffractive imaging: applications and limitations," Opt. Express **11**, 2344–2353 (2003).
5. J. Miao, D. Sayre, and H. N. Chapman, "Phase retrieval from the magnitude of the Fourier transforms of nonperiodic objects," J. Opt. Soc. Am. A **15**, 1662–1669 (1998).
6. I. Robinson, I. Vartanyants, G. Williams, M.A.Pfeifer, and J. Pitney, "Reconstruction of the shapes of gold nanocrystals using coherent x-ray diffraction," Phy. Rev. Lett. **87**(19), 195505 (2001).
7. J. Miao, T. Ishkawa, B. Johnson, E. H. Anderson, B. Lai, and K. O. Ohodgson, "High resoluton 3D x-ray diffraction microscopy," Phys. Rev. Lett. **89**, 088302 (2002).
8. J. C. H. Spence, U. Weierstall, and M. Howells, "Phase recovery and lensless imaging by iterative methods in optical x-ray and electron diffraction," PPhilos. Trans. R. Soc. London, Ser. A **360**, 875–895 (2002).
9. Y. Nishino, J. Miao, and T. Ishikawa, "Image reconstruction of nanostructured nonperiodic objects only from oversampled hard x-ray diffraction intensities," Phys. Rev. B **68**, 220101 (2003).

#142626 - $15.00 USD Received 15 Feb 2011; revised 7 May 2011; accepted 25 May 2011; published 1 Jun 2011

10. S. Marchesini, H. He, H. N. Chapman, S. Hau-Riege, A. Noy, M. R. Howells, U. Weierstall, and J. C. H. Spence, "X-ray image reconstruction from a diffraction pattern alone," Phy. Rev. B **68**, 140101 (2003).
11. G. Williams, M. A. Pfeifer, I. Vartanyants, and I. Robinson, "Three-dimensional imaging of microstructure in au nanocrystals," Phys. Rev. Lett. **90**, 195505 (2003).
12. G. Huldt, A. Szoke, and J. Hajdu, "Diffraction imaging of single particles and biomolecules," J. Struct. Biol. **144**, 219–227 (2003).
13. S. Hau-Riege, H. Szoke, H. N. Chapman, A. Szoke, S. Marchesini, A. Noy, H. He, M. Howells, U. Weierstall, and J. C. H. Spence, "Speden: reconstructing single particles from their diffraction patterns," Acta Crystallogr. Sect. A: Found. Crystallogr. **60**, 294–305 (2004).
14. D. Shapir, P. Thibault, T. Beetz, V. Elser, M. Howells, C. Jacobsen, J. Kirz, E. Lima, H. Miao, A. M. Neiman, and D. Sayre, "Biological imaging by soft x-ray diffraction microscopy," Proc. Natl. Acad. Sci. U.S.A. **102**(43), 15343–15346 (2005).
15. H. N. Chapman, A. Barty, S. Marchesini, A. Noy, S. P. Hau-Riege, C. Cui, M. R. Howells, R. Rosen, H. He, J. C. H. Spence, U. Weierstall, T. Beetz, C. Jacobsen, and D. Shapiro, "High-resolution ab initio three-dimensional x-ray diffraction microscopy," J. Opt. Soc. Am. A **23**, 1179–1200 (2006).
16. H. N. Chapman, S. Bajt, A. Barty, W. H. Benner, M. J. Bogan, M. Frank, S. P. Hau-Riege, R. A. London, S. Marchesini, E. Spiller, A. Szke, and B. W. Woods,"Ultrafast coherent diffraction imaging with x-ray free-electron lasers," Proc. FEL, 805–811 (2006).
17. J. M. Rodenburg, A. C. Hurst, A. G. Cullis, B. R. Dobson, F. Pfeiffer, O. Bunk, C. David, K. Jefimovs, and I. Johnson, "Hard x-ray lensless imaging of extended objects," Phy. Rev. Lett. **98**, 034801 (2007).
18. C. Song, H. Jiang, A. Mancuso, B. Amirbekian, L. Peng, R. Sun, S. S. Shah, Z. H. Zhou, T. Ishikawa, and J. Miao, "Quantitative imaging of single, unstained viruses with coherent x rays," Phys. Rev. Lett. **101**, 158101 (2008).
19. M. M. Seibert, T. Ekeberg, F. R. N. C. Maia, M. Svenda, J. Andreasson, O. Jonsson, D. Odic, B. Iwan, A. Rocker, D. Westphal, M. Hantke, D. P. DePonte, A. Barty, J. Schulz, L. Gumprecht, N. Coppola, A. Aquila, M. Liang, T. A. White, A. Martin, C. Caleman, S. Stern, C. Abergel, V. Seltzer, and J.-M. Claverie, "Single mimivirus particles intercepted and imaged with an x-ray laser," Nature **470**, 78–81 (2011).
20. G. Bortel and G. Faigel, "Classification of continuous diffraction patterns: a numerical study," J. Struct. Biol. **158**, 10–18 (2007).
21. N.-T. D. Loh and V. Elser, "Reconstruction algorithm for single-particle diffraction imaging experiments," Phy. Rev. E **80**, 026705 (2009).
22. R. Fung, V. Shneerson, D. K. Saldin, and A. Ourmazd, "Structure from fleeting illumination of faint spinning objects in flight," Nat. Phys. **5**, 64–67 (2009).
23. D. K. Saldin, V. L. Shneerson, R. Fung, and A. Ourmazd, "Structure of isolated biomolecules obtained from ultrashort x-ray pulses: exploiting the symmetry of random orientations," J. Phys.: Condens. Matter **21**, 134014 (2009).
24. M. van Heel, "Angular reconstruction: a *posteriori* assignment of projection directions for 3d reconstruction," Ultramicroscopy **21**, 111–124 (1987).
25. N. A. Farrow and F. P. Ottensmeyer, "A *posteriori* determination of relative projection directions of arbitrarily oriented macromolecules," J. Opt. Soc. Am. A **9**, 1749–1760 (1992).
26. M. van Heel, "Single-particle electron cryo-microscopy: towards atomic resolution," Q. Rev. Biophys. **33**, 307–369 (2000).
27. R. Miles, "On random rotations in r^3," *Biometrika* **52**, 636–639 (1965).
28. J. Lipfert, D. Herschlag, and S. Doniach, "Riboswitch conformations revealed by small-angle x-ray scattering," Methods Mol. Biol. **540**, 141–159 (2009).

1. Introduction

With the advent of the X-ray free electron laser, it is possible to image a single biomolecule through coherent diffraction imaging [1–4], albeit at lower wavelengths. Many promising approaches have been suggested for accomplishing this goal [5–18]. In a typical coherent X-ray imaging experiment, single biomolecules are injected into a pulsed X-ray beam. The elastically scattered photons are recorded by a detector array placed downstream. Due to the ultrashort time scale of the pulse, the scattering pattern can be recorded before the molecule is destroyed [19]. The recorded scattering data are then sorted according to the orientations of the biomolecules. Signal strength can be improved by averaging shots of the biomolecules having the same orientation, which will benefit subsequent steps of phase retrieval and image reconstruction. Due to the small size of the biomolecule and small elastic scattering cross-section of the electron, we encounter two problems for the coherent X-ray diffraction imaging. One is that the detected signal will be mostly dominated by Poisson noise. This is especially the case when

#142626 - $15.00 USD Received 15 Feb 2011; revised 7 May 2011; accepted 25 May 2011; published 1 Jun 2011

the scattering angle is large, as the scattering is mostly in the forward direction for coherent illumination. The other difficulty is that the biomolecule itself is intrinsically a 3D structure and its orientation is not controlled. Thus, orientation classification is the first crucial step of data processing in coherent X-ray imaging.

Early work on classification was based on large scattering angle data where the diffraction pattern is modeled as a speckle pattern [12, 20]. In practice, this is primarily true for large-angle scattering where the detected photon level is very low with the current or near future X-ray photon budgets. Thus the classification schemes based on large-angle-scattering-data are not currently practical. Recently there are two interesting developments where Expectation-Maximization algorithms is utilized for classification [21, 22]. The signal level for successful classifications is much reduced. Another imaging scheme is to simultaneously inject many randomly oriented biomolecules [23]. The scattered photons are much more and the orientation classification is unnecessary in this setup.

In this paper we focus on scattering by a single particle and propose a classification scheme based on patterns of low scattering angles where a substantial number of photons can be detected. Further the comparison among different rotations is not directly based on the pattern itself but on compressed signals which are the radial and azimuthal distributions of the scattering pattern. As both of these distributions are signals integrated over many pixels, the variance in the radial or azimuthal distribution is much smaller than that in the individual pixels. Consequentially a classification can be relatively immune to the shot noise and the results of the classification can be trustworthy. An example of utilizing those compressed signals is given in this paper for classification of in-plane rotations (i.e., relative orientations between biomolecules are within a plane parallel to the detector plane).

2. Weak elastic scattering process and the issue of shot noise

In a typical setup for coherent diffraction imaging of single biomolecules, the samples are injected into the waist of the focused coherent X-ray beam and a detector is placed downstream to record the scattering pattern. The scattered pattern recorded by the detector array, $I(x,y)$, can be written as

$$I(x,y) = \Phi_{inc} r_e^2 \Omega_{pix} \left| \sum_{n=1}^{N_{atom}} f_n(x'_n, y'_n, z'_n) exp(j2\pi \frac{xx'_n + yy'_n}{\lambda z}) \right|^2 , \quad (1)$$

where r_e is the free electron scattering length, Φ_{inc} is the incident fluence and Ω_{pix} is the solid angle of one detector pixel extended with respect to the biomolecule. The coordinates and atomic form factor of the n^{th} atom in the biomolecule are described by (x'_n, y'_n, z'_n) and f_n, respectively. The pixel coordinates for the detector array are (x,y,z) and the wavelength of the X-ray is *lambda*. The center region of the detector array has a hole to pass the unscattered photons. In this paper we are interested in the small-angle scattering patterns, the variation of the atomic form factor over the scattering angles can thus be ignored and f_n is assumed to be a real function. The DNA polymerase delta (PDB ID:3IAY, molecular weight: 113822; total number of atoms: $\sim$ 8.2K) is used as an example of a typical biomolecule to calculate the diffraction pattern and perform orientation classifications.

In this paper we also assume that the noise is mainly shot noise, the detected photon number in any pixel is a Poisson distributed random variable whose mean value can be calculated from Eq. (1). In Fig. 1 we show a typical scattering pattern by the single biomolecule when the incident photon number is infinite (noise free), 2×10^{14} and 2×10^{12} per pulse. The beam is assumed to be 100 nm and the X-ray wavelength is 1.5$\AA$. The detector array is 150 mm away from the sample and the pixel size is 0.11 mm.

#142626 - $15.00 USD Received 15 Feb 2011; revised 7 May 2011; accepted 25 May 2011; published 1 Jun 2011

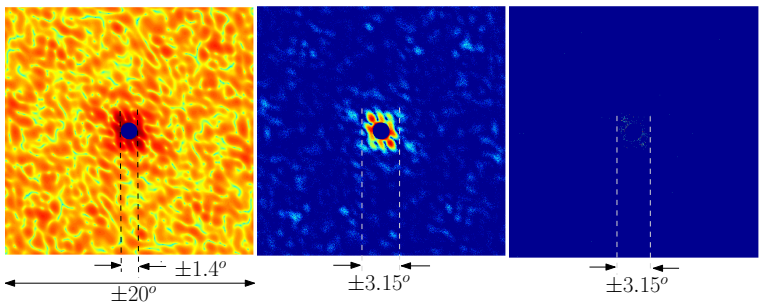
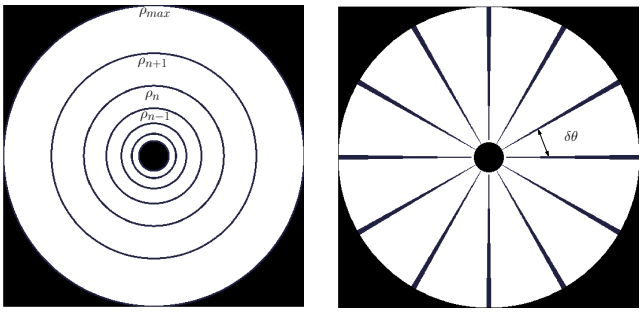


Fig. 1. Typical simulated scattering patterns of a DNA polymerase delta (PDB ID:3IAY). The data are shown in logarithmic scale. The radius of the hole in the center of the detector array is 3.72 mm. Left: "noise free" pattern; Middle: incident photon number is 2×10^{14} and the number of collected photons is 818875; Right: incident photon number is 2×10^{12} and the number of collected photons is 8114.

From Fig. 1 we can see that for the "noise free" case, the scattering pattern looks like speckles in regions corresponding to large scattering angles. However, with a finite incident photon number such as 2×10^{14} photons per pulse, the contrast of the speckles is diminished and some even disappear. For the 2×10^{12} photons per pulse case, which is the current available level at LCLS, the number of photons scattered with scattering angles larger than 2° is so little that basically all speckles disappear. It will be very difficult to distinguish any exposures from each other, not to mention conducting orientation classification with this incident photon level.

Consider an annular area of the scattering pattern with radius $\rho_{min}\leq\rho\leq\rho_{max}$. We divide this area into regions of radial or azimuthal segments as shown in Fig. 2(a) and 2(b). The boundaries



(a) (b)

Fig. 2. Detector array is divided into radial (a) or azimuthal segments (b).

between these radial segments are rings with radii $\rho_1,\rho_2,...,\rho_N$, which can be defined as

$$\rho_{n+1}=\rho_n+\Delta\rho_n,\quad n=0,1,2,...N-1; \tag{2}$$
$$\Delta\rho_{n+1}=\left(1+\left(\frac{\Delta\rho_n}{\rho_n}\right)^2\right)\Delta\rho_n,$$

where $\rho_0=3.72mm,\Delta\rho_0=0.66mm$. We see that the width of the outer ring is larger than the inner rings. This will help to compensate for the decreasing signal strength with increasing

#142626 - $15.00 USD Received 15 Feb 2011; revised 7 May 2011; accepted 25 May 2011; published 1 Jun 2011

radius. The radial distribution, $I_\rho(n)$, can be written as,

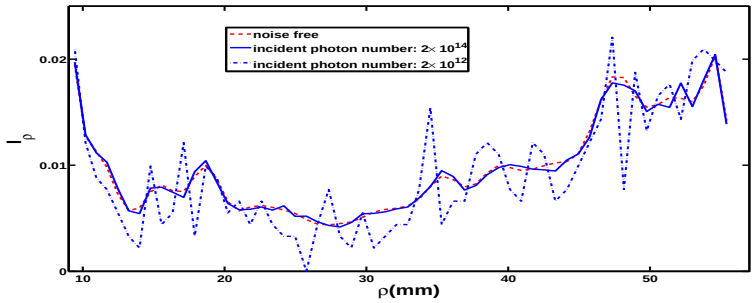
$$I_\rho(n) = \frac{\int_{\rho_n}^{\rho_{n+1}} \int_0^{2\pi} I(\rho,\theta) d\theta \rho d\rho}{\int_{\rho_{min}}^{\rho_{max}} \int_0^{2\pi} I(\rho,\theta) d\theta \rho d\rho}, \quad n = 0,1,2,...N-1 \tag{3}$$

Note that I_ρ is normalized with respect to the total scattered photons in the effective detection area. In this paper we choose $N = 60$, which is able to significantly reduce the variance in the radial distribution and still give us sufficient sensitivity to rotations as we will see in the following parts of this paper.

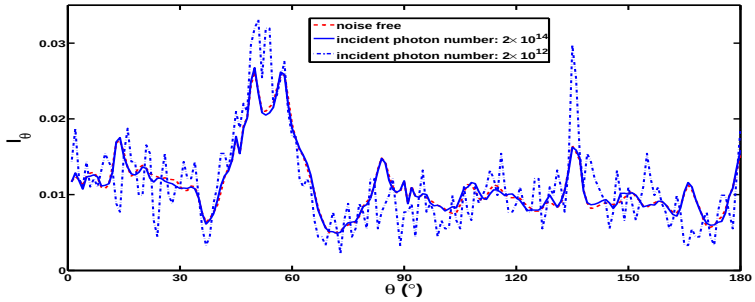
In the simulation study we choose $\delta\theta = 1^\circ$. As the scattering pattern is symmetric about the origin, we limit the value of θ to be within $[0, 180^\circ]$. The azimuthal distribution can be written as:

$$I_\theta(n) = \frac{\int_{\theta_n}^{\theta_{n+1}} \int_{\rho_{min}}^{\rho_{max}} I(\rho,\theta) \rho d\rho d\theta}{\int_0^{\pi} \int_{\rho_{min}}^{\rho_{max}} I(\rho,\theta) \rho d\rho d\theta}, \quad n = 0,1,2,...179. \tag{4}$$

In order to study the effect of the noise, the mean value of the radial and azimuthal distributions are first calculated from the "noise free" diffraction patterns. The noisy versions of the radial or azimuthal distribution are calculated from the noisy diffractions when the incident photon number is 2×10^{14} and 2×10^{12} respectively. A typical radial and azimuthal distributions and their mean values are shown in Fig. 3(a) and 3(b) respectively. We see that when the



(a)



(b)

Fig. 3. Typical radical (a) and azimuthal (b) distributions. Three cases are studied: the mean value case (dotted line); the 2×10^{14} photon number case (solid line) and the 2×10^{12} photon number case (dashed line).

incident photon number is 2×10^{14}, the noisy version of the radial and azimuthal distribution

#142626 - $15.00 USD Received 15 Feb 2011; revised 7 May 2011; accepted 25 May 2011; published 1 Jun 2011

is similar to its mean value. However when the incident photon number is 2×10^{12}, the noisy version of the radial and azimuthal distribution deviates substantially from their corresponding mean values, which will affect the reliability of the classification adversely.

In Figs. 4 we show surface plots when the biomolecule rotates about the x, y, and z axes. We see that with rotations about x or y, both the radial and the azimuthal distributions change smoothly with rotation angles. For rotations about z, the radial distribution stays the same while the azimuthal distribution shifts continuously with the rotation angle. Thus rotation about z (also called an in-plane rotation) can be recognized by studying the evolution of the radial distributions, and the rotation angles can be identified by finding the amount of the shift of the azimuthal distribution. For rotations other than along the z-axis, both the radial and azimuthal distributions are going to change smoothly with the rotation angles. Thus a distance between two distributions, e.g. Euclidean distance, can be a measure describing the global changes during continuous rotation. A small distance between distributions means that the two distributions are quite similar to each other and the relative rotation angle between them is small. In a separate simulation we modified a large biomolecule (PDB ID: 3I55, size $\sim 200\mathring{A}$) by removing all atoms greater than $90\mathring{A}$ from the centroid and showed that the same process can be easily applied to spherical molecules.

3. Euler angles and the possibility of in-plane rotation

In the previous sections we have devised a method to preprocess the scattering data to improve the data consistency for the classification of orientations. In this section, we will use introduce Euler angles, which have been widely used in the cryo-EM field, to represent arbitrary orientations [24–26].

An arbitrary rotation of an angle ϕ about an arbitrary axis can be described by three consecutive rotations, i.e.,

$$R(\phi) = R_z(\gamma)R_y(\beta)R_z(\alpha); \qquad -\pi \leq \alpha, \gamma \leq \pi, \quad 0 \leq \beta \leq \pi, \tag{5}$$

where R is the rotation matrix that depends on the rotation axis and rotation angle ϕ. The first rotation is about the intrinsic z-axis, the second rotation is about the intrinsic y-axis and the last one is again a rotation about the intrinsic z-axis.

As the low-angle scattering pattern is proportional to the square of the amplitude of the Fourier transform of the scattering potential, which is a real and even function, we can limit the ranges of α, β and γ to $[0, \pi]$.

The last γ-rotation is the so-called in-plane rotation because the z-axis of the intrinsic coordinate and the z-axis of the laboratory coordinates are co-axial. If we can classify the in-plane rotation out of all possible rotations and know the angle of the in-plane rotation of a selected pattern with respect to a reference pattern, we can rotate that pattern to overlap with the reference pattern. This will give us an opportunity to increase the SNR by averaging over all the in-plane rotations. The extent of the SNR improvement depends on the number of in-plane rotations.

Consider the random orientation of a single biomolecule when injected into the passage of the X-ray beam, the probability of the orientation which can be described by the Euler angles (α, β, γ) and can be written as [27]

$$p(\alpha, \beta, \gamma) = \frac{\sin\beta}{8\pi^2}. \tag{6}$$

Due to the symmetry of the scattering pattern, the effective ranges for Euler angles are within $[0, \pi]$, thus we can write the effective probability density function in terms of $\alpha, \cos\beta, \gamma$ as

$$p_{eff}(\alpha, \cos\beta, \gamma) = \frac{1}{2\pi^2}, \qquad 0 \leq \alpha, \beta, \gamma \leq \pi. \tag{7}$$

#142626 - $15.00 USD Received 15 Feb 2011; revised 7 May 2011; accepted 25 May 2011; published 1 Jun 2011

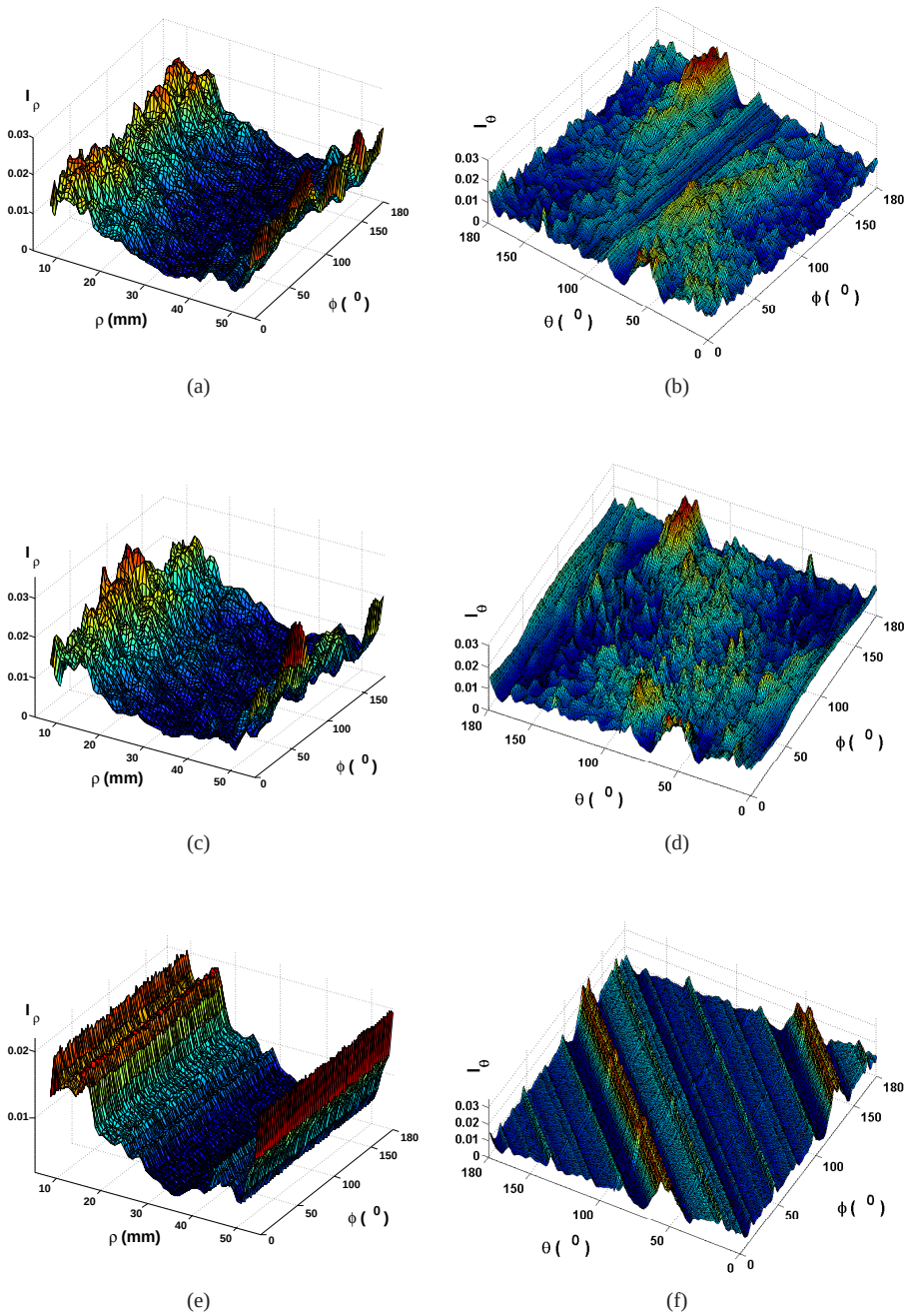


Fig. 4. Surface plots of the radial (left column) and azimuthal (right column) distributions of the scattering patterns when the biomolecule rotates about the x-axis (a,b), y-axis (c,d), and z-axis (e-f). The incident photon number is 2×10^{14}.

#142626 - $15.00 USD Received 15 Feb 2011; revised 7 May 2011; accepted 25 May 2011; published 1 Jun 2011

Thus the probability density function is uniform within a rectangular volume of $2\pi^2$ with Euler rotations defined by α, γ and $\cos\beta$. With a total number N_{total} of random rotations, the number density for α and γ is $N_{total}^{1/3}/pi$ and $N_{total}^{1/3}/2$ for $\cos\beta$. Thus the mean number of in-plane rotations is expected to be:

$$N_{inplane} = \frac{{N_{total}}^{1/3}}{n}. \tag{8}$$

where n is the number of classes for each Euler angles. Thus the SNR of averaging over all the in-plane rotations can be further improved by $\frac{N_{total}^{1/6}}{\sqrt{n}}$ times compared to averaging over only same orientations.

4. In-plane rotation recognition through the radial distribution

We can also see from the surface plots shown in Fig. 4(e) that the radial distribution stays unchanged with continuous rotation about z. Thus, we may identify the in-plane rotation by studying the radial distribution.

Consider a distance defined as

$$D_\rho(I_{ref}, I_{test}) = \left| \sum_{n=1}^{N} (I_\rho^{ref}(n) - I_\rho^{test}(n))^2 \right|^{1/2}, \tag{9}$$

where I_{test} and I_{ref} are the scattering pattern before and after the in-plane rotation, respectively, I_ρ^{ref} and I_ρ^{test} are the radial distributions calculated according to Eq. (3). This "radial" distance should be zero if there is no shot noise, but with shot noise unavoidable, this distance should still be minimum if it is to be an effective classification measure. Thus we start from the biomolecule in an initial orientation, compute the scattering pattern according to Eq. (1), include the Poisson distributed noise and compute the corresponding radial distribution. This radial distribution is taken as the reference distribution I_ρ^{ref}. Then we rotate the biomolecule along y- and z-axis by angles from 0 to 180°. The corresponding noisy scattering pattern I_{test} and radial distribution I_ρ^{test} are computed in the same way as I_{ref} and I_ρ^{ref}. Then the distance between I_ρ^{test} and the reference distribution I_ρ^{ref} is computed according to Eq. (9). Clearly this distance is a function of the rotation of the test pattern with respect to the reference pattern. We repeat the distance calculation 100 times independently and calculate the mean and variance of the distance from those results (Fig. 5). In order to understand the effect of the incident photon number on the distance calculation, we have used two incident photon number levels: 2×10^{14} (Fig. 5(a)) and 2×10^{12} (Fig. 5(b)). We see that with both cases, the mean value of the distance maintains the lowest level for the in-plane rotations and increases with the out-of-plane rotation (rotation about y-axis) in general. Zoomed-in versions of the distance curves around the reference position are plotted in Fig. 5(c) and 5(d). We can see that for the case of 2×10^{14} incident photon number, the variance in the distance is so small that we can differentiate even a half degree of out-of-plane rotation from in-plane rotations with high certainty. This means that the sensitivity of the above classification is about 0.5°. With less incident photon number, the scattering patterns are more noisy and have a larger variance (Fig. 5(d)). The sensitivity of the classification is about $\pm 2.5^\circ$ in this case.

5. Finding the in-plane rotation angles through the azimuthal distribution

After having identified scattering patterns corresponding to in-plane rotations, we can find the angle of the in-plane rotation by studying the azimuthal distribution I_θ. This angle information

#142626 - $15.00 USD Received 15 Feb 2011; revised 7 May 2011; accepted 25 May 2011; published 1 Jun 2011

is needed if we hope to average over in-plane rotations. In this section we describe a simple method to find this angle.

If I_θ^{ref} is the azimuthal distribution of the reference pattern and I_θ^{test} is the azimuthal distribution of the test pattern with only an in-plane rotation γ with respect to the reference pattern, then these two distributions should be shifted with respect to each other. This can be clearly seen in the surface plot shown in Fig. 4(f). The amount of the shift should be the angle of the in-plane rotation of the biomolecule with respect to its original position.

Consider a new function, I'_θ, which is a circularly-shifted version of I_θ^{test}, i.e.,

$$I'_\theta(\theta,\theta_{shift}) = I_\theta^{test}(\theta-\theta_{shift}), \tag{10}$$

where θ_{shift} is the shift angle. When the shift angle equals the in-plane rotation angle γ, the shifted azimuthal distribution I'_θ shall overlap with the reference azimuthal distribution I_θ^{ref}. Thus we can define a distance between I'_θ and I_θ^{ref} as:

$$D_\theta(I_{ref},I_{test},\theta_{shift}) = \left|\int_0^\pi \left(I_\theta^{ref} - I_\theta^{test}(\theta-\theta_{shift})\right)^2 d\theta\right|^{1/2}, \tag{11}$$

where I_{ref} and I_{test} are the scattering patterns before and after the in-plane rotation, respectively. This distance is a function of the shift angle θ_{shift}. When the shift angle is exactly the in-plane rotation angle, this distance should be zero if there is no shot noise. With shot noise being unavoidable, this distance should reach the minimum when $\theta_{shift} = \gamma$. For example, when the in-plane rotation is 90 degrees, we compute the distance according to Eq. (11) as a function of the shift angle θ_{shift} . In order to study the effect of the incident photon number on the distance, we have again studied two cases, 2×10^{14} and 2×10^{12} (Fig. 6). We can see that the distance reaches the minimum when the shift angle equals the in-plane rotation angle for both incident photon number levels. Note the steepness of the curves around the in-plane rotation angle, indicating that the distance is quite sensitive to the shift angle. In this example, it is about 0.5°, which is much smaller than the sensitivity for identifying the in-plane rotations from all possible rotations. Thus in the next few sections, we will use the sensitivity of the radial distance as the limit of the overall sensitivity of in-plane rotation classification.

With in-plane rotations classified (section 5) and the angle of in-plane rotation determined as described in this section, we can rotate those in-plane rotated patterns to overlap the reference patterns, and average those patterns to improve the signal-to-noise ratio. In Fig. 7 we show a typical single shot of the scattering pattern and the averaged pattern from 19 shots with different in-plane rotations. Comparing these two images we can clearly see that the signal-to-noise ratio is improved after the averaging, especially in regions away from the center.

6. Accuracy of the classification vs photon budget

As we saw in the previous sections, the simple distances can be used to compare the radial or azimuthal distributions to distinguish between in-plane rotations and out-of-plane rotations, and to find the degree of in-plane rotation. In this section we are going to study the performance of the classification scheme for different incident photon numbers. To be specific, we want to know the ratio of the correct classification results from the results of the proposed classification method.

In order to visualize the success rate of the proposed algorithm, we have generated a number of random orientations and their corresponding diffraction patterns, then used the method described in Section 5 to find the in-plane rotations. The Euler angles for those orientations are limited to $\{0^o, 180^o\}$. Twenty random values for each $\alpha, \cos\beta, \gamma$ are generated independently.

#142626 - $15.00 USD Received 15 Feb 2011; revised 7 May 2011; accepted 25 May 2011; published 1 Jun 2011

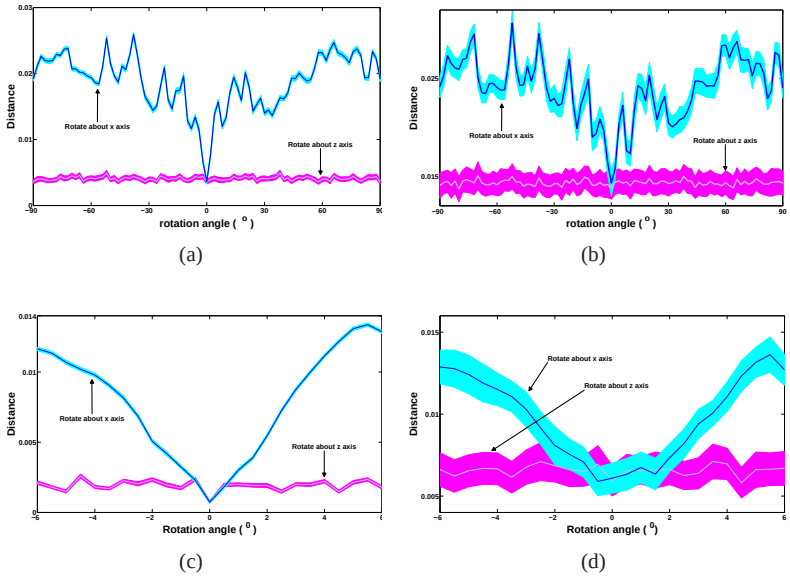


Fig. 5. The radial distance as a function of the rotation angle when the protein rotates about y, and z respectively. The curve for rotation along z is shown in pink; in blue for rotation along y. Two incident photon levels are studied here: (a) 10^{14}; (b) 10^{12}; The mean values of the distances are shown in the solid lines; The color-shaded areas represent $\pm\sigma$ around the mean values for both the rotation about y and z respectively. σ is the variance. (c)-(d) are zoomed-in versions of (a) and (b) respectively.

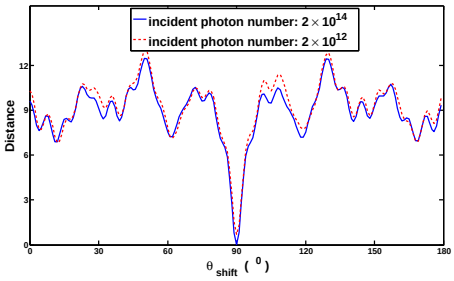
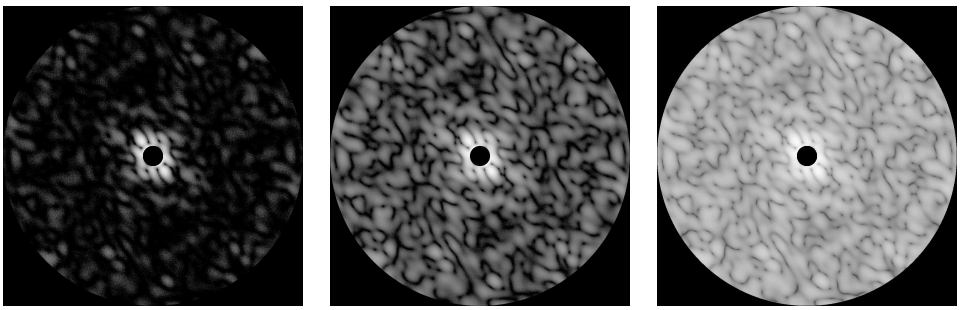


Fig. 6. Azimuthal distance reaches minimum at the in-plane rotation angle. Two incident photon number levels are studied: 2×10^{14}(solid line) and 2×10^{12}(dashed line).

#142626 - $15.00 USD Received 15 Feb 2011; revised 7 May 2011; accepted 25 May 2011; published 1 Jun 2011



(a) (b) (c)

Fig. 7. (a) Scattering pattern from a single measurement; (b) Scattering pattern after averaging over 19 patterns with different in-plane rotations; (c) "Noise free" pattern.

Thus a total number of orientations is $20 \times 20 \times 20 = 8000$, ignoring the inconsequential fact that the ranges for those three random variables $(\alpha, \cos\beta, \gamma)$ are not all equal to each other. The diffraction patterns for the protein with those orientations are generated using Eq. (1). Then we select an arbitrary pattern as a reference frame and assign the rest of the patterns to be the test patterns. The corresponding radial distribution for the reference frame and the test patterns are computed according to Eq. (3). Then the distance is calculated using Eq. (9) to compare the reference and test frames.

Consider a total number N_{total} of patterns from randomly oriented biomolecules, the total number of pairwise distances, N_d, will be $N_{total} \times (N_{total} - 1)$ if each pattern is compared to other patterns in the data set. Note that the comparison between any two patterns are performed twice since each one has the opportunity to be a reference pattern. We also assume that the number of random values for each Euler angle is $N_{total}^{1/3}$. Thus the percentage for in-plane rotations among all possible rotations shall be $N_{total}^{1/3}/N_{total}$. The total number of distances that describe in-plane rotations, N_{in}^d, is expected to be

$$N_{in}^d = N_{total} \times (N_{total} - 1) \times \frac{N_{total}^{1/3}}{N_{total}}. \tag{12}$$

As the distances describing the in-plane rotations should be zero if there is no noise in the experiment, but with noise, we choose N_{in}^d shortest distances as arising from comparison of two patterns who differ by at most an in-plane rotation. Of course, the noise inherent in the frames will tend to mix the in-plane rotations and out-of-plane rotations, especially those patterns with small out-of-plane rotations.

As we have generated the patterns and know the orientations of the test frames, we can verify the results of the classification with the known information. Thus we can define the accuracy of the classification as:

$$\eta = \frac{N_{true}}{N_{classified}}, \tag{13}$$

where N_{true} and $N_{classified}$ are the number of true and assigned in-plane rotations, respectively. This accuracy rate can provide a quantitative measure of the classification performance. For a certain classification scheme, it depends mainly on two factors, noise level and the size of the orientation classes. For patterns that were found to be in-plane rotations, their values for α and β should be identical in the ideal case, but with noise, we say the classification is accurate if α and β differ by an amount smaller than the size of the orientation class. In this section we will

#142626 - $15.00 USD Received 15 Feb 2011; revised 7 May 2011; accepted 25 May 2011; published 1 Jun 2011

test the class size from 1° to 5°. In Table 1 we list the accuracy of these simulation results for different incident photon numbers and different orientation class sizes.

Table 1. Accuracy of the In-Plane Rotation Classification[a]

	Incident photon number		
Size of the orientation class	2×10^{14}	2×10^{13}	2×10^{12}
1°	98.9% (84.9%)	97.6% (78.0%)	79.3% (49.3%)
2°	99.5% (93.6%)	98.9% (90.3%)	86.2% (73.3%)
3°	99.5% (95.9%)	99.4% (94.5%)	92.1% (82.2%)
4°	99.5% (96.9%)	99.4% (95.9%)	95.0% (85.9%)
5°	99.5% (97.2%)	99.4% (96.8%)	97.3% (90.5%)

[a]The values outside the parenthesis are for a complete data set and those inside the parenthesis are for the smaller data set described in the text.

We can see that even with the current incident level, the success of the classification can be as high as 79% with a class size of 1°, which is required for atomic resolution. If we can relax the requirement on the sensitivity of the classification, we can achieve 92% accuracy for the class size of 3° and 97% accuracy for 5°. In this way we can break the data set into smaller groups and use more sophisticated algorithms such as described in [21] or [22] to find additional information. Research in this direction is underway. With higher photon levels, almost 99% accuracy can be achieved for class size of 1°. The accuracy from a mere guess is expected to be $N_{inplane}/N_{total}$, which is only 0.25% in this case.

Note that with the limited number of generated random orientations due to available computing speed, the orientations which are close to each other in the R^3 space may not have been fully simulated in the above example. Thus we perform another round of simulations where the Euler angles are limited to $\alpha \in (0,10^\circ), \cos\beta \in (0.9,1)$. As the radial distributions are invariant to in-plane rotations, we choose the limit on the in-plane rotation angle still to be $(0,180^\circ)$. For each Euler angle we generate 10 random numbers resulting in 1000 patterns. With a much smaller range of Euler angles, the percentage for two orientations whose Euler angles $(\alpha_i,\beta_i, i=1,2)$ are within 1° becomes much higher. In our simulation, the percentage is about 48%, i.e., almost half of the orientations have a similar orientation. The accuracy in this small data set can be lower compared to the entire data set. In Table 1 we list the accuracy (numbers inside the parenthesis) for different photon numbers and orientation class sizes.

We see that even in this smaller data set where almost half of the orientations have a nearby orientation, the success of the classification is still almost 50%. Note that with a smaller data set, the expected success rate by a mere guess is bigger. In the case we have described above, it is about $10/1000 = 1\%$ and so the proposed classification scheme still gives a much better result.

7. Discussion and summary

In this paper we have shown a method to compress the 2-dimensional diffraction patterns into 1-dimensional distributions, namely, radial and azimuthal distributions. With the compressed signals, the variance is greatly reduced. Those signals can be utilized as a first step in processing the data from coherent X-ray experiments. As an example, we have used a simple distance measure computed from the radial and azimuthal distributions to compare the test frame and the reference frames. As the radial distributions are invariant to in-plane rotations, we can use the radial distribution to find the in-plane rotations. The data is divided into classes of (α,β). With

#142626 - $15.00 USD Received 15 Feb 2011; revised 7 May 2011; accepted 25 May 2011; published 1 Jun 2011

the proposed classification scheme we can achieve 79% accuracy even with current incident photon number. With a protein that is bigger than used in this paper or possesses additional symmetries, a good classification may be achieved with an even lower incident photon level. We have performed two simulations to test the performance of the proposed classification scheme. The performance for a real situation should be between these two cases.

With the degree of in-plane rotation identified, we can rotate one pattern to overlap with the corresponding reference pattern. This provides us an opportunity to be able to average over more patterns improving the signal-to-noise level further compared to averaging over only patterns with the same orientations. The method is non-iterative in nature, which makes it a good candidate for data preprocessing.

Another interesting aspect of the suggested compression method is that the radial distribution is invariant to in-plane rotations. This property could be used in conjunction with the manifold method described in [22]. A hyper-space constructed from the radial distribution shall exhibit a 2-manifold depending only on α and β. With fewer parameters to estimate, searching the manifold will be easier, which could potentially lower the required photon level even more. Further as the variance in the radial distributions is much smaller than those in the individual pixels, the searching of the 2-manifold will be more robust to noise, thus offering us a possibility to work with even fewer photons per pixel.

As far as atomic resolution is concerned, it is associated with large-angle scattering data, which lies on the Ewald sphere. As the signal-to-noise ratio is extremely low in these regions, averaging diffraction patterns within the same orientation class is necessary to obtain meaningful large-angle-scattering data. If the size of the orientation class is large, information will be lost during averaging. For a biomolecule of size a, the maximum displacement of the same atom within one orientation class of size $\delta\theta$ can be estimated as

$$|\delta\mathbf{r}| = \frac{a}{2}\delta\theta.$$

In order to maintain atomic resolution, the displacement of atoms within one class should be smaller than an angstrom. Thus for a biomolecule of size about 10 *nm*, such as the one considered in this paper, the size of the orientation class should be around 1.2°. From Table 1 we see that the classification scheme proposed in this paper can divide the data into orientation classes with size of 1°, thus maintaining the atomic resolution desired for coherent X-ray imaging.

In this paper, the biomolecules are assumed to be rigid during the injection and measuring time. Other than their orientations and relative positions in the X-ray beam, the biomolecules are also assumed to be identical, i.e. no shot-to-shot variation of the structure or conformation is considered. Different conformations of the molecules will result in different diffraction patterns. The small-angle scattering data and the resulting radial and azimuthal distributions will encode information on different confirmations [28]. The classification scheme proposed in this paper will hence distinguish different classes to different conformations. Whether a certain class belongs to an out-of-plane orientation of the reference molecule in the same conformation or an orientation class of the molecule in a different conformation can be addressed in the next step of the classification. One possible method for such purpose is the manifold method [22]. A general heterogeneous sample may not be solely classified by means of the low-angle scattering data, large-angle-scattering may have to be additionally included in the classification algorithm. This is a research direction we are currently pursuing.

A common technique for solving the orientation problem in cryo-EM is the common line approach [24–26]. Because of the extremely low signal-to-noise level, this method is not applicable in current x-ray diffraction experiments. Current classification work on X-ray diffraction data [21, 22] is primarily focused on the direct 2D diffraction patterns. We propose to compress the diffraction data by radial and azimuthal integrals, so that the orientation information can

#142626 - $15.00 USD Received 15 Feb 2011; revised 7 May 2011; accepted 25 May 2011; published 1 Jun 2011

still be extracted, but the variance is greatly reduced. As we have shown in Section 2, we can use the simple difference measure as described in Section 4 and 5 for classification of in-plane rotations. We can also use more elaborate methods for a complete orientation classification, such as [21, 22], with the advantage of improved SNR.

In summary, the radial and azimuthal distributions of scattering patterns from single biomolecules has shown a great success in classifying in-plane rotations by using a simple distance measure. With more sophisticated algorithms such as the manifold method, we may be able to perform a complete orientation classification and produce the expected diffraction patterns with enough sensitivity to obtain atomic resolution.

Acknowledgment

This work is funded by UCOP LAB FEE. This work is also performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under the contract DE-AC52-07NA27344.

#142626 - $15.00 USD Received 15 Feb 2011; revised 7 May 2011; accepted 25 May 2011; published 1 Jun 2011