Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

An improved integration method in serial femtosecond crystallography

Kun Qu, Liang Zhou and Yu-Hui Dong*

Beijing Synchrotron Radiation Facility, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, People's Republic of China

Correspondence e-mail: dongyh@ihep.ac.cn

Received 30 March 2013
Accepted 28 January 2014

Recent experiments in serial femtosecond crystallography (SFX) have demonstrated the feasibility of obtaining structural information from nanoscale crystals using X-ray free-electron lasers (XFELs). However, millions of crystals are required to determine one reliable structure. Here, an improved integration algorithm for SFX data processing is reported. By evaluating the dimensions of each crystal and correcting for the geometric factors of single patterns, the effective diffraction intensities, as opposed to the directly measured single-shot pattern diffraction intensities, can be merged to acquire more accurate integrated intensities which can be used for structure determination. This improvement enhances the quality of electron-density maps and decreases the number of diffraction patterns that are needed to solve the crystal structure in SFX experiments.

1. Introduction

The advent of serial femtosecond crystallography (SFX) represents a novel approach for structure determination of macromolecules, aided by new technologies such as continuous liquid micro-jets (DePonte *et al.*, 2008) and new detector devices (Strüder *et al.*, 2010; Philipp *et al.*, 2011). Compared with conventional crystallography, SFX does not require large crystals. It is believed to be a potential solution for challenging structure determinations of difficult-to-crystallize molecules and even membrane proteins in the near future (Johansson *et al.*, 2012; Koopmann *et al.*, 2012). Therefore, improving SFX data-processing methods is of great interest to the crystallography community, as these methods are necessary for the wide application of SFX in the structural biology field.

The general procedure for X-ray diffraction data processing comprises several steps, which include auto-indexing, intensity integration and scaling, with several intermediate refinement cycles (Rossmann & van Beek, 1999). A number of successful methods have been described to perform intensity integration of diffraction data in traditional protein crystallography (Otwinowski & Minor, 1997; Kabsch, 1988; Leslie, 1999). Nevertheless, because of the size distribution and random orientation among the crystals used, existing integration methods cannot be applied to SFX. To calculate the integrated intensities from a series of single-shot patterns from huge numbers of crystals with different sizes and random orientations, a Monte Carlo integration method (Kirian *et al.*, 2010) was first introduced to process SFX data. Because the crystal hit rate is very low, and successful Monte Carlo integration requires massive amounts of data, the experimental data collection has to last for up to several weeks in order to obtain

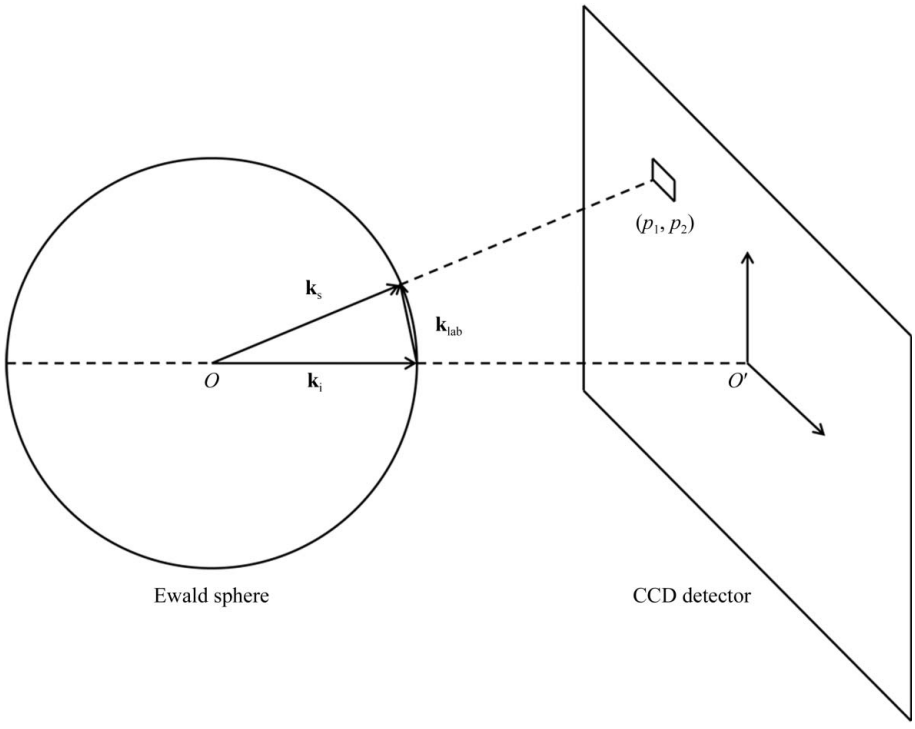


Figure 1
A schematic of the general SFX experimental setup in the laboratory coordinate frame. $\mathbf{k}_s$ and $\mathbf{k}_i$ refer to the scattered and incident vectors, respectively. The scattering vector $\mathbf{k}_{lab} = \mathbf{k}_s - \mathbf{k}_i$. (p_1, p_2) refers to the coordinates of a pixel on the detector. The corresponding relations between pixels and scattering vectors can be established.

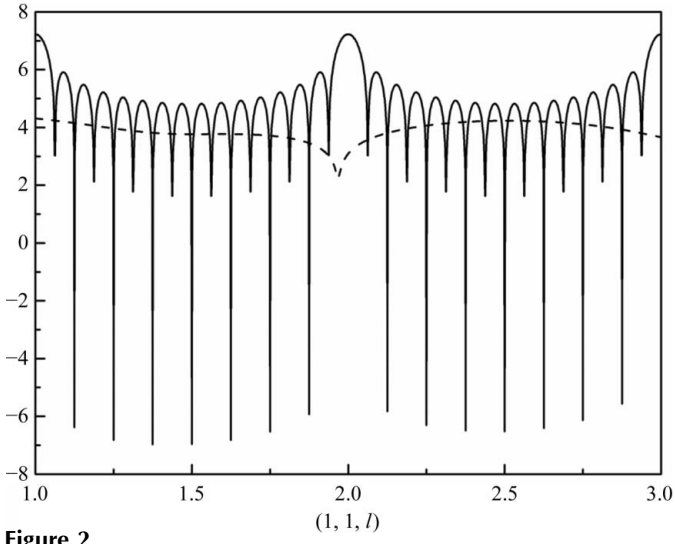


Figure 2
The logarithmic scale of $|G_n(\mathbf{k})|^2$ (solid line) and $|F(\mathbf{k})|^2$ (dashed line) as a function of the reciprocal-lattice vector (1, 1, l). (Reproduced with permission from Zhou *et al.*, 2013).

interpretable electron-density maps. In a recent experiment (Boutet *et al.*, 2012), the structure determination of hen egg-white lysozyme at 1.9 Å resolution was performed from 12 247 and 10 575 individual indexed diffraction patterns using 40 and 5 fs X-ray pluses, respectively, at the Linac Coherent Light Source.

In order to decrease the required number of patterns, one promising solution is to reduce the number of random variables in the Monte Carlo integration. Using the existing auto-indexing algorithm (Powell, 1999) for determining crystal orientations, we propose a search algorithm for estimating the sizes of each crystal. By correcting for the geometric factors related to the estimated crystal sizes, the effective diffraction intensities can be derived from the observed diffraction spot intensities in the pattern, and these scaled diffraction intensities can be merged over all snapshot patterns. By eliminating the influence of the crystal size distribution, the accuracy of the integrated intensities is improved and the number of patterns needed to obtain reliable structures is decreased.

2. Methods

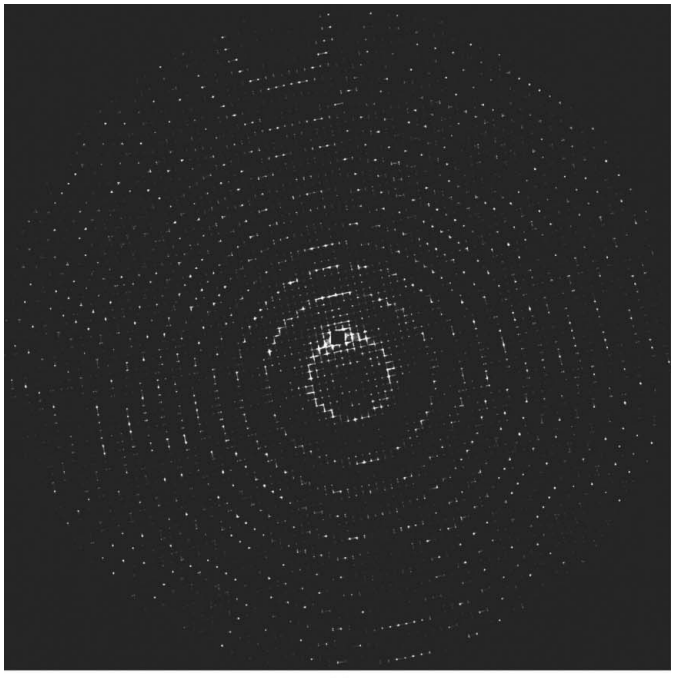
Protein crystals are relatively weak scatterers. In the traditional protein crystallography experiment, a crystal size of 0.2–0.4 mm is regarded as optimal to meet the general requirements (Drenth, 2007). The crystal is continuously rotated and a series of diffraction patterns are collected. Diffraction spot intensities from different patterns are then integrated and scaled with a linear weight factor (Otwinowski & Minor, 1997), because the crystal does not change during data collection. With ultrashort high-intensity X-ray pulses from free-electron lasers, even nanometre-sized crystals can provide sufficient diffraction signals in SFX (Chapman *et al.*, 2011). However, one crystal can only survive long enough to give only one pattern because of the destructive intensity of XFEL. Therefore, data sets obtained in SFX experiments consist of many single-shot patterns from different crystals. Fig. 1 shows a schematic of the general SFX experimental setup. The expression for ideal snapshot diffraction intensity in SFX is given by

$$I_n(\mathbf{k}) = I_0 r_e^2 |F(\mathbf{k})|^2 |G_n(\mathbf{k})|^2 \Delta\Omega_n(\mathbf{k}), \quad (1)$$

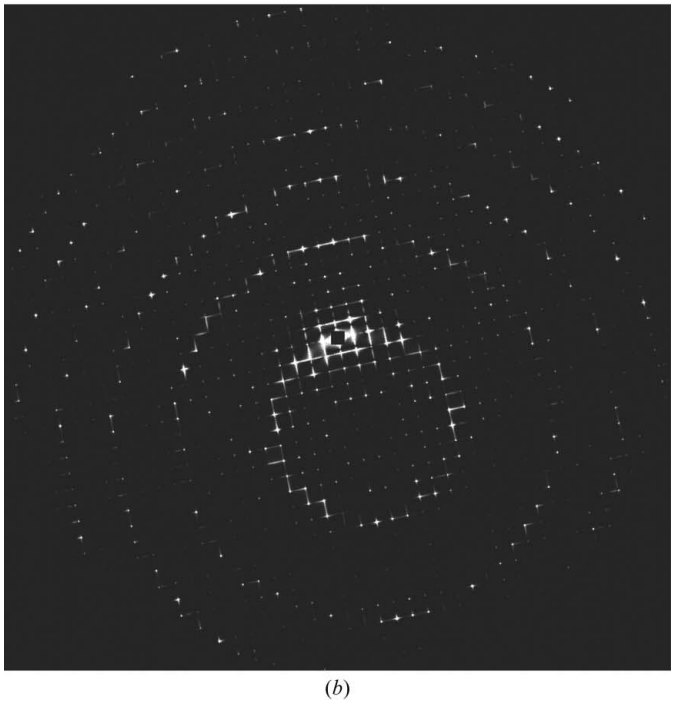
where I_0 is the incident photon flux density (photons s^{-1} m^{-2}) and r_e is the classical Thomson scattering length. $F(\mathbf{k})$ is the structure factor, which is defined in terms of the unit cell and does not differ between crystals. $\Delta\Omega_n(\mathbf{k})$ is the solid angle subtended by a detector pixel along the scattered vector $\mathbf{k}_s$. $|G_n(\mathbf{k})|^2$ is the geometric factor of the nth crystal defined as

$$|G_n(\mathbf{k})|^2 = \left[\frac{\sin(\pi h N_{n,x})}{\sin(\pi h)}\right]^2 \left[\frac{\sin(\pi k N_{n,y})}{\sin(\pi k)}\right]^2 \left[\frac{\sin(\pi l N_{n,z})}{\sin(\pi l)}\right]^2, \quad (2)$$

where $\mathbf{k}$ is the reciprocal-lattice vector and (h, k, l) are the values of the scattering vector projected on the three crystal axes in reciprocal space. In units of unit cells, the size of the nth crystal is $N_n = N_{n,x} \times N_{n,x} \times N_{n,z}$. If $\mathbf{k}_0$ denotes the reciprocal-lattice vector with integer Miller indices, $|G_n(\mathbf{k}_0)|^2$ is equal to N_n^2 under the assumption of an ideal crystal.



(*a*)



(*b*)

Figure 3
Typical diffraction patterns of a perfect nanocrystal with 13 × 12 × 13 unit cells along the *a*, *b* and *c* axes at 2.5 Å (*a*) and 5.0 Å (*b*) resolution simulated without Poisson noise added. Non-Bragg mid-peaks between neighbouring Bragg peaks caused by the geometric factor are obvious in both situations.

The geometric factor values vary greatly, whereas the squares of the structure-factor amplitudes are more uniform in reciprocal space, as illustrated in Fig. 2. For very small crystals, Bragg peak broadening is observed, and non-Bragg mid-peaks can be distinguished between neighbouring Bragg peaks (Chapman *et al.*, 2011). Typical diffraction patterns simulated according to (1) at high and low resolution are illustrated in Fig. 3. Non-Bragg mid-peaks between neighbouring Bragg peaks are obvious in both situations. In addition, Bragg peak broadening may also result from crystal mosaicity, wavelength dispersion, beam divergence and other factors. Here, we limit the discussion to the influence of the geometric factor on the diffraction intensity.

It should be noted that the scattering vector $\mathbf{k}$ is defined above in the crystal coordinate frame. In Fig. 1, the scattered vector $\mathbf{k}_s$ can be calculated from the coordinates of each detector pixel according to the experimental geometry, and the scattering vector $\mathbf{k}_{lab} = \mathbf{k}_s - \mathbf{k}_i$ in the laboratory coordinate frame can be obtained. In the meantime, the orientation of each crystal is first determined in the indexing step. Following X-ray diffraction theory (Rossmann & van Beek, 1999), the orientation matrix can be defined as

$$\mathbf{M} = \begin{pmatrix} a_x^* & b_x^* & c_x^* \\ a_y^* & b_y^* & c_y^* \\ a_z^* & b_z^* & c_z^* \end{pmatrix}, \tag{3}$$

where a_i^*, b_i^*, and c_i^* ($i = x, y, z$) are projections of the reciprocal unit-cell axes in the laboratory coordinate frame. The scattering vector $\mathbf{k}$ in the crystal coordinate frame is calculated by

$$\mathbf{k} = M^{-1}\mathbf{k}_{lab} \tag{4}$$

and the geometric factor at $\mathbf{k}$ is determined by (2). We can then define the average of the product of $|G_n(\mathbf{k})|^2$ and $\Delta\Omega_n(\mathbf{k})$ over a pixel (p_1, p_2) as

$$I_{G,n}(p_1, p_2) = \langle |G_n(\mathbf{k})|^2 \Delta\Omega_n(\mathbf{k}) \rangle_{(p_1,p_2)}. \tag{5}$$

In a typical experiment, since the pixel dimensions are far smaller than the sample-to-detector distance, (5) may be simplified as

$$I_{G,n}(p_1, p_2) = \Delta\Omega_n(p_1, p_2) \langle |G_n(\mathbf{k})|^2 \rangle_{(p_1,p_2)} \tag{6}$$

and the pixel intensity $I_n(p_1, p_2)$ can be treated as a function of $I_{G,n}(p_1, p_2)$ and $|F(\mathbf{k})|^2$. For nanocrystals, both Bragg peaks and non-Bragg mid-peaks are more apt to broaden resulting from more visible reflection shape transforms in reciprocal space. Fig. 2 and (1) show that the fluctuation of $I_n(p_1, p_2)$ between adjacent pixels principally depends on $I_{G,n}(p_1, p_2)$. By comparing the mean difference MD between $I_n(p_1, p_2)$ and $I_{G,n}(p_1, p_2)$ on adjacent pixels, a G-search algorithm (Zhou *et al.*, 2013) is rewritten to analyze the peak profiles as

$$\mathrm{MD} = \frac{1}{\sum_{(p_1,p_2)}\left(1+\sum_{m=-1}^{1}1\right)} \times \sum_{(p_1,p_2)}\left\{\left|\left|\frac{I_n(p_1+1,p_2)}{I_n(p_1,p_2)}\right| - \left|\frac{I_{G,n}(p_1+1,p_2)}{I_{G,n}(p_1,p_2)}\right|\right| + \sum_{m=-1}^{1}\left|\left|\frac{I_n(p_1+m,p_2+1)}{I_n(p_1,p_2)}\right| - \left|\frac{I_{G,n}(p_1+m,p_2+1)}{I_{G,n}(p_1,p_2)}\right|\right|\right\}, \quad (7)$$

where p_1, p_2 are the coordinates of a pixel on the detector and only those pixels with observed intensities are taken into account to avoid numerical instability. An exhaustive search over a certain range of $N_{n,x}$, $N_{n,y}$ and $N_{n,z}$ can locate the global minimum of the MD and provide feedback on an estimated crystal size.

Therefore, the influence of the geometric factors can be eliminated using the following algorithm:

$$I_n^{\mathrm{eff}}(\mathbf{k}) = \begin{cases} \dfrac{I_n(\mathbf{k})}{I_{G,n}(\mathbf{k})} & |\mathbf{k}-\mathbf{k}_0| \le \delta_n \\ \dfrac{I_n(\mathbf{k})}{N_n^2\Delta\Omega(\mathbf{k})} & |\mathbf{k}-\mathbf{k}_0| > \delta_n \end{cases}. \quad (8)$$

In this equation, $I_n^{\mathrm{eff}}(\mathbf{k})$ is the effective diffraction intensity from a unit cell at the reciprocal-lattice vector $\mathbf{k}$ and is proportional to the square of the structure-factor amplitude. δ_n is defined to specify the extent of Bragg peak broadening at $\mathbf{k}_0$ in the nth pattern. Because the full-width at half-maximum (FWHM) of the main peak in $|G_n(\mathbf{k})|^2$ is inversely proportional to the crystal size N_n (Fig. 2), δ_n can be quantified as

$$\delta_{n,x} = \frac{w}{2N_{n,x}}, \quad \delta_{n,y} = \frac{w}{2N_{n,y}}, \quad \delta_{n,z} = \frac{w}{2N_{n,z}}, \quad (9)$$

where w is a weight factor that is used to evaluate the influence of other factors on the broadening of the diffraction spot. In an ideal situation, w = 2.0 indicates the first zero of the geometric factor. Theoretically, δ_n varies with different crystals because the profiles of reciprocal-lattice points are variable during the data integration. In the existing Monte Carlo integration method, the geometric factor is treated as a mean value for all patterns and the integrated intensities are calculated by averaging the directly measured intensities in all patterns (Kirian *et al.*, 2010). In general, if the number of patterns is large enough, the obtained integrated intensities will be sufficiently accurate.

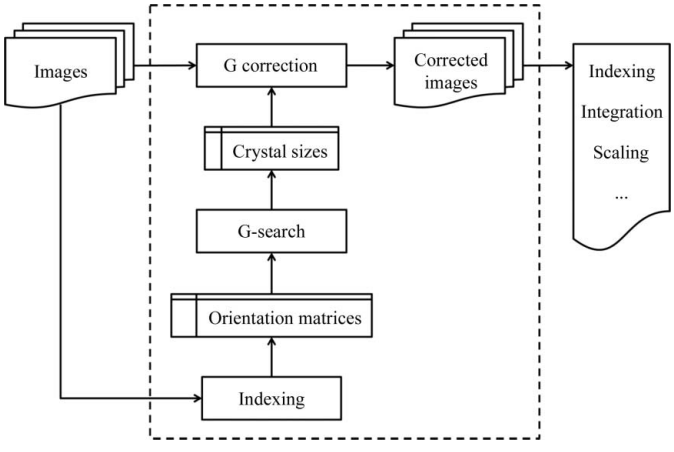


Figure 4
Data-flow diagram of the geometric factor correction algorithm. The multiple steps in the dotted box are those added before the traditional procedures.

Table 1
Simulations of the incident X-ray free-electron laser.

Wavelength (Å)	1.5
Radius of X-ray focus (μm)	2.0
Pulse fluence at sample (photons per pulse)	5.0×10^{12}

A significant innovation in our algorithm is that the effective intensities, instead of the directly measured intensities, are taken into account in the integration step. Therefore, the randomness from crystal sizes is eliminated in SFX. Fig. 4

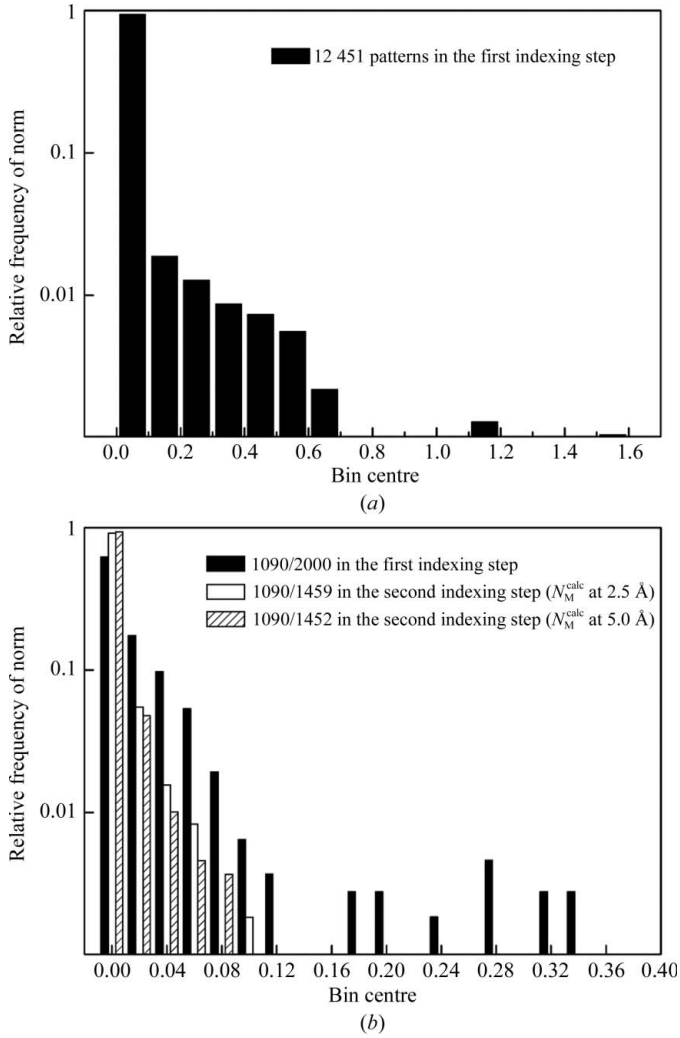


Figure 5
Frequency histogram of the matrix norm defined in (10). The relative frequency is on a logarithmic scale. (*a*) The matrix norm of the 12 451 indexed patterns in the first indexing step. (*b*) The matrix norm of the 1090 common patterns successfully indexed in all three cases: the first 2000 patterns of the 12 451 patterns in the first indexing step, the 1459 corrected patterns with $N_{\mathrm{M}}^{\mathrm{calc}}$ at 2.5 Å resolution and the 1452 corrected patterns with $N_{\mathrm{M}}^{\mathrm{calc}}$ at 5.0 Å resolution in the second indexing step.

shows a data-flow diagram of the geometric factor correction algorithm. The multiple steps in the dotted box are those added before the traditional procedures of indexing, integration and scaling. Once the individual images have been selected and indexed, the crystal orientation matrices are determined. In the G-search algorithm, these approximate orientation matrices are employed to calculate the reciprocal-lattice vectors of the corresponding pixels with observed intensities. With estimated crystal sizes as input, the correction of the geometric factors for different patterns according to (8) is completed as a pre-processing step and the improved integration algorithm of the effective intensity is implemented. The whole process can be described as follows.

(i) An individual image is first indexed and the crystal orientation matrix is determined. Based on the experimental setup, the scattering vector $\mathbf{k}_{\rm lab}$ in the laboratory coordinate

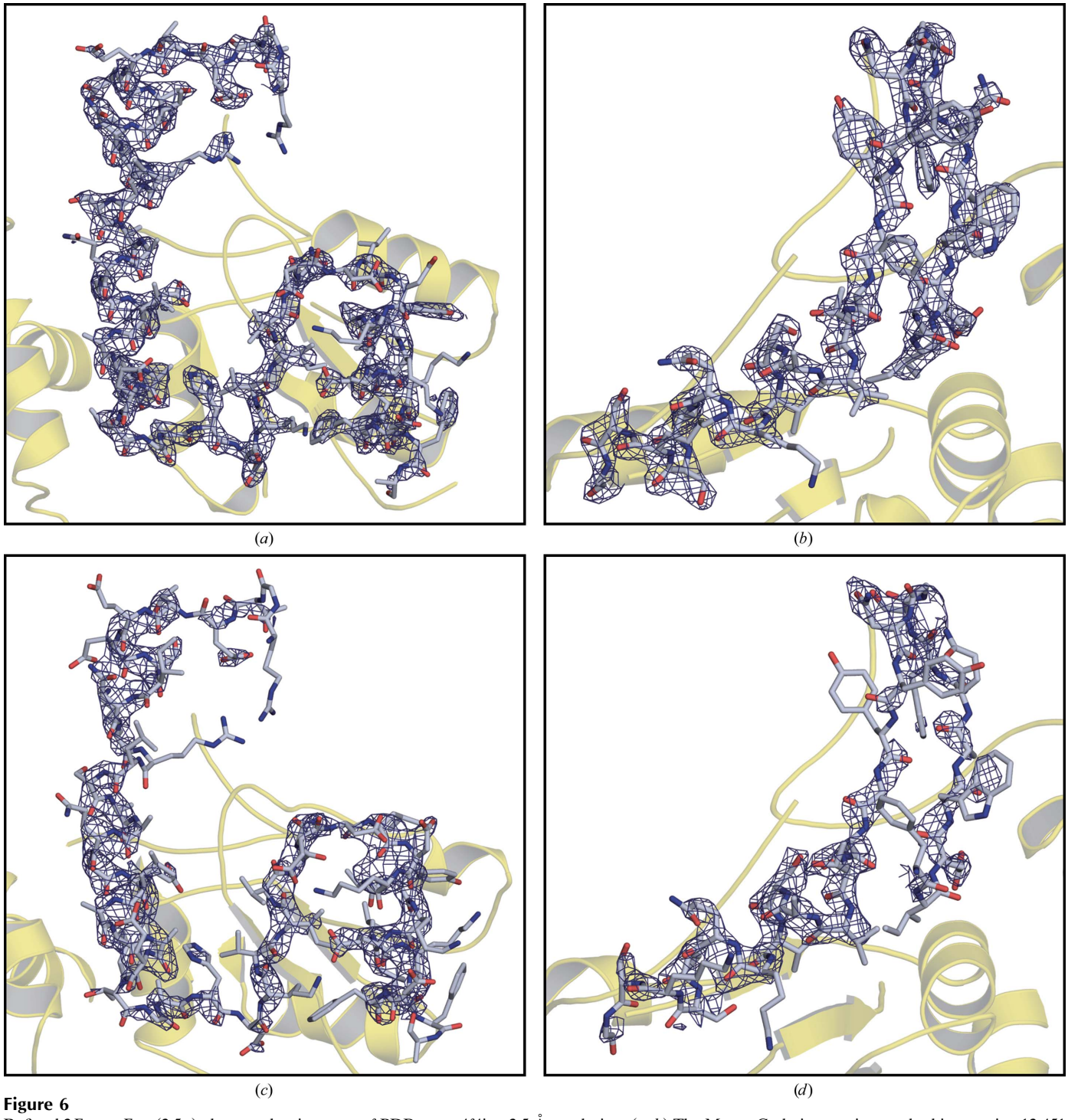


Figure 6
Refined $2F_{\rm obs} - F_{\rm calc}$ (2.5σ) electron-density maps of PDB entry 4f4j at 2.5 Å resolution. (*a*, *b*) The Monte Carlo integration method integrating 12 451 simulated patterns. (*c*, *d*) The Monte Carlo integration method integrating 2000 simulated patterns.

Table 2
Estimated crystal sizes of the first 20 patterns and that (the 74 695th) which showed the maximum deviation. Results were calculated by the G-search algorithm using simulated patterns at 5.0 and 2.5 Å resolution, respectively. The subscripts M and Q denote the orientation matrices determined by the *MOSFLM* program and the exact matrices.

No.	$N^{\rm exact}$	$N_{\rm M}^{\rm calc}$ (5.0 Å)	$N_{\rm Q}^{\rm calc}$ (5.0 Å)	$N_{\rm M}^{\rm calc}$ (2.5 Å)	$N_{\rm Q}^{\rm calc}$ (2.5 Å)	norm
1	(24, 16, 18)	(16, 23, 21)	(24, 16, 18)	(9, 12, 12)	(12, 18, 12)	0.002262
3	(23, 12, 18)	(9, 10, 20)	(23, 12, 22)	(10, 10, 14)	(23, 18, 22)	0.070935
20	(21, 24, 20)	(21, 24, 20)	(21, 24, 20)	(9, 12, 10)	(17, 25, 17)	0.001812
30	(27, 23, 19)	(27, 26, 21)	(27, 23, 20)	(23, 13, 13)	(23, 13, 11)	0.001485
39	(16, 17, 17)	(9, 9, 7)	(16, 17, 17)	(9, 9, 8)	(16, 12, 10)	0.120248
54	(22, 21, 20)	(21, 10, 22)	(22, 21, 21)	(13, 28, 16)	(22, 12, 17)	0.002256
58	(20, 25, 21)	(20, 25, 21)	(20, 25, 21)	(26, 25, 16)	(26, 20, 15)	0.001059
59	(14, 24, 17)	(24, 14, 17)	(14, 24, 20)	(11, 14, 10)	(15, 12, 10)	0.002995
61	(15, 15, 8)	(28, 9, 7)	(15, 15, 8)	(17, 15, 9)	(14, 17, 8)	0.161747
81	(25, 26, 18)	(13, 13, 18)	(25, 26, 18)	(18, 13, 12)	(20, 13, 12)	0.000740
92	(9, 18, 9)	(9, 9, 9)	(9, 18, 9)	(9, 10, 8)	(9, 10, 10)	0.051976
101	(14, 25, 14)	(25, 14, 7)	(14, 26, 14)	(28, 16, 17)	(15, 26, 18)	0.014520
103	(19, 22, 13)	(24, 22, 13)	(20, 22, 13)	(14, 19, 17)	(19, 14, 16)	0.004430
122	(20, 25, 18)	(28, 20, 18)	(20, 25, 18)	(21, 28, 12)	(26, 19, 12)	0.002232
125	(26, 22, 20)	(9, 10, 7)	(26, 22, 20)	(9, 12, 9)	(13, 16, 16)	0.038014
131	(27, 27, 18)	(9, 9, 8)	(27, 28, 18)	(16, 27, 20)	(28, 24, 14)	0.003555
144	(21, 19, 15)	(19, 21, 18)	(21, 19, 15)	(14, 26, 9)	(10, 14, 9)	0.003406
160	(21, 23, 15)	(23, 21, 15)	(21, 23, 15)	(14, 24, 9)	(25, 15, 9)	0.002801
167	(16, 26, 16)	(16, 9, 16)	(16, 26, 16)	(16, 28, 21)	(16, 26, 22)	0.002942
171	(26, 23, 19)	(26, 23, 19)	(26, 23, 19)	(24, 20, 13)	(24, 22, 13)	0.001549
74695	(10, 14, 8)	(9, 25, 7)	(10, 14, 8)	(12, 16, 9)	(10, 16, 8)	1.664295

Table 3
Convergence of the G-correction algorithm. $N_{\rm M}^{\rm calc}$ are the estimated crystal sizes reduced from 2.5 or 5.0 Å resolution patterns with the G-search algorithm. w is the weight factor in (9). The success rates of the second indexing step in the four cases are listed. $R_{\rm split}$ is defined by (11).

$N_{\rm M}^{\rm calc}$ (Å)	w	Indexing rate	$R_{\rm split}$
2.5	1.0	1459/2000	0.2469
5.0	1.0	1452/2000	0.2336
2.5	1.6	1559/2000	0.4236
5.0	1.6	1637/2000	0.2909

frame can be calculated for every pixel on the detector. The scattering vector **k** in the crystal coordinate frame is then found using (4).

(ii) For each pattern, the estimated crystal size is determined by an exhaustive search.

(iii) The geometric factors of images are corrected for to obtain the effective intensities.

(iv) The corrected images are then used in the subsequent indexing step.

3. Results and discussion

3.1. Simulation of data and convergence of the G-search algorithm

For the data presented here, the detector model consisted of 1456 × 1456 110 µm pixels. Diffraction patterns were simulated with the *CrystFEL* software v.0.4.0 (White *et al.*, 2012). In this simulation, a single CPU in a 3.10 GHz quad-core processor was used, a typical SFX detector geometry was assumed and Poisson noise was added. Other parameters in the simulated experiments are listed in Table 1.

We chose entry 4f4j (Johnston *et al.*, 2011; Table 4) in the Protein Data Bank, the unit-cell size of which was large enough to produce obvious geometric factor effects, as the protein model. Crystal orientations were randomly generated and subjected to a uniform distribution. The numbers of unit cells N_x, N_y and N_z in one nanocrystal were normally distributed between values of 10 and 30, corresponding to crystal sizes of 100–300 nm. Two sets of data were simulated at 2.5 and 5.0 Å resolution at the side of the detector with sample-to-detector distances of 11.46 and 25.77 cm, respectively. The 2.5 Å high-resolution data set contained 100 000 patterns. Firstly, these patterns were imported into the *CrystFEL* software, and about 12.45% of them were indexed successfully. The low indexing rate can be attributed to the small number of photons per pulse at the sample and the weak diffractive ability of small protein crystals. A control group of 2000 patterns at 5.0 Å resolution were then simulated. The crystal sizes and orientations of these low-resolution patterns were input as the same as those of the first 2000 high-resolution patterns of the 12 451 indexed patterns. The high-resolution data set was used to determine the structure and the low-resolution data set was used to test the robustness of the G-search algorithm with orientation errors and different resolutions.

In the first indexing step for the high-resolution data set, the orientation matrices of individual crystals were determined from the *MOSFLM* software (Leslie, 1992). To estimate errors, the matrix norm can be defined as

$$\text{norm} = \|\mathbf{M}^{-1}\mathbf{Q} - \mathbf{I}\|_2, \tag{10}$$

where **M** is the orientation matrix determined by the software, **Q** is the exact orientation matrix generated from the data simulation and **I** is a unit matrix. The resulting statistical values of norm for 12 451 images ranged from a minimum value of 1.79×10^{-4} to a maximum of 1.67. Fig. 5(*a*) shows that significant errors exist for a few crystals with regard to indexing single patterns in SFX. The G-search subroutines were then used to evaluate the 2000 crystal sizes with 2.5 and 5.0 Å resolution patterns. For comparison, the determined and exact orientation matrices were used, respectively. The results of the first 20 patterns and that which showed the maximum deviation are listed in Table 2. Considering the equivalence innate to the symmetry and the indistinguishability of the *a* and *b* axes in the $P4_32_12$ space group, the numbers of unit cells along the two axes are commutable.

Search results demonstrate that the errors in the crystal orientations have certain influences on the G-search algorithm. Differences between $N_{\rm M}^{\rm calc}$ and $N_{\rm Q}^{\rm calc}$, for either 5.0 or 2.5 Å resolution, are negligible most of the time, except for

Table 4
Data-simulation and refinement statistics.

Values in parentheses are for the highest resolution shell. N.A., not applicable.

Parameter	Monte Carlo integration	Improved integration		
Space group	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$
Unit-cell parameters (Å)	$a = b = 103.682$, $c = 130.881$	$a = b = 103.682$, $c = 130.881$	$a = b = 103.682$, $c = 130.881$	$a = b = 103.682$, $c = 130.881$
No. of integrated patterns	12451 [100000]	2000	1459 [N_M^{calc} at 2.5 Å]	1452 [N_M^{calc} at 5.0 Å]
No. of reflections	N.A.	N.A.	N.A.	N.A.
No. of unique reflections	24249	20115	21150	20784
Resolution limits (Å)	63.96–2.50 (2.60–2.50)	63.96–2.50 (2.63–2.50)	81.27–2.50 (2.63–2.50)	63.96–2.50 (2.63–2.50)
Completeness† (%)	95.65 (86.80)	79.34 (52.30)	83.42 (61.30)	81.99 (59.40)
R_{split}	0.0864	0.1809	0.2469	0.2336
Average B factor† (Å^2)	52.00	55.50	54.00	59.60
R_{work}/R_{free}†	0.1955/0.2472	0.3143/0.4142	0.2216/0.2897	0.2322/0.2935
R.m.s.d., bonds† (Å)	0.0090	0.0113	0.0109	0.0089
R.m.s.d., angles† (°)	1.177	1.631	1.308	1.174

† Calculated with *MolProbity* (Chen *et al.*, 2010).

some particular cases. For example, at high resolution the calculated values of N_M^{calc} and N_Q^{calc} are similar for the pattern displaying maximal deviation. However, at low resolution the crystal sizes estimated from the G-search algorithm are strangely inconsistent. Another example is the 20th pattern. Even though the error in the crystal orientation is close to zero (norm = 1.812×10^{-3}), the results at 2.5 and 5.0 Å disagree with each other. Thus, crystal orientation errors are unlikely to account for the faults in searching the crystal dimensions.

It is worth mentioning that upon examination of N^{exact} and N_M^{calc} at 2.5 Å resolution in Table 2, the reduced crystal sizes from the high-resolution patterns are very different from the exact sizes. At high resolution, because the signal-to-noise ratio is low and the diffraction peak profiles are sharp, the region with observed intensities around single diffraction spots on the detector only contributes to a few pixels. This poor sampling of diffraction peaks can result in a great loss of the geometric factor information. In other words, all of the information on the undulation of the diffraction intensities within a single pixel is destroyed, leaving only a digitally recorded intensity summation. The dimensions of the detector pixel cause the extra peak broadening. After considering machinery and equipment factors, the crystal sizes are apparently underestimated at high resolution by (7). However, we will see below that these underestimated crystal sizes in the G-search algorithm could be more effective to correct for the geometric factors of single patterns.

3.2. Results of the Monte Carlo integration method

After the first indexing step, the program *process_hkl* based on the Monte Carlo integration method (Kirian *et al.*, 2010) was applied to merge the selected 12 451 patterns, of which the first 2000 were processed alone as a control. Initial phases of the structure factors were obtained by molecular replacement with the program *Phaser* (McCoy *et al.*, 2007), using the main chains of PDB entry 4f4j as a search model. Iterative rounds of model building and restrained refinement were performed by *PHENIX* (Adams *et al.*, 2010) and *Coot* (Emsley & Cowtan, 2004). Some snapshots of the refined structures are shown in Fig. 6. Using the Monte Carlo integration method, there is a quite spectacular contrast in that the electron density in Figs. 6(*a*) and 6(*b*) is interpretable by integrating 12 451 patterns, while the electron density of the side chains, and even the main chains, is universally absent (Figs. 6*c* and 6*d*) when there is a shortage of images (only 2000 patterns). The Monte Carlo integration method closely relies on the number of patterns.

3.3. Results of the improved integration algorithm with G-correction algorithm

The geometric factors were calculated with the estimated crystal sizes. The G-correction algorithm was then performed as shown in Fig. 4 and the directly measured intensities were divided by the geometric factors. Inevitably, the effective diffraction intensities for each pixel were lower than the corresponding initial intensities. Before the second indexing step, the corrected patterns were multiplied by an empirical weight factor of about 10.0. In this way, the data completeness of the final integrated intensities can be retained from the Monte Carlo integration method to our improved integration method.

When w equals 1.0 in (9), we corrected for the geometric factors of the first 2000 indexed images at 2.5 Å resolution with the estimated crystal sizes obtained from the 2.5 Å resolution images. These corrected images were then indexed by the *CrystFEL* software again. Only 1459 patterns were indexed successfully. As a control, we also corrected the same 2000 images at 2.5 Å resolution with the estimated crystal sizes obtained from the 5.0 Å resolution images and 1452 patterns were indexed successfully. We calculated the error of the determined orientation matrix in the second indexing step. Statistical results of the 1090 common patterns successfully indexed in all the three cases are compared in Fig. 5(*b*). We found that the indexing accuracy after the G-correction processing with N_M^{calc} at both 2.5 and 5.0 Å resolution is higher than the first indexing step. However, the drop in success rate is undesirable in the second indexing step. By comparison, we found that the maximum effective intensities in some corrected patterns are nearly 100 times lower than the

maximum directly measured intensities in the initial patterns, and this might be the reason for the drop in the indexing rate.

To evaluate the data quality after the G-correction processing, $R_{\rm split}$ (White *et al.*, 2012) was calculated with even/odd diffraction intensities as

$$R_{\rm split} = \frac{1}{2^{1/2}} \frac{\sum_{hkl} |I_{\rm even} - I_{\rm odd}|}{\frac{1}{2}\sum_{hkl}(I_{\rm even} + I_{\rm odd})}. \quad (11)$$

The $R_{\rm split}$ factors of the G-correction algorithm were 0.2336 with the crystal sizes calculated from the 5.0 Å resolution patterns and 0.2469 with the crystal sizes calculated from the 2.5 Å resolution patterns. To explain the influence of w on $R_{\rm split}$, the G-correction step was performed for a second time with w = 1.6 and $R_{\rm split}$ factors were acquired in both cases. As shown in Table 3, $R_{\rm split}$ in the high-resolution situation is obviously higher. Such a difference between $R_{\rm split}$ factors at low and high resolution is reasonable because the G-search algorithm is more unstable at high resolution and the G-correction algorithm has to settle with larger errors. When the value of w is increasing, the difference is more visible. In addition, a larger w seems to be conducive for increasing the success rate in the second indexing step. However, it should be treated with caution because weak intensities of satellite peaks around the main reflection are possibly fixed to a high value and merged into the Bragg peak intensities by the *MOSFLM* software. A smaller w in (9) at high resolution helps to reduce the higher $R_{\rm split}$ factor and to improve the integral precision.

The electron density was resolved by molecular replacement following the same procedures as the Monte Carlo integration method. It can be seen in Fig. 7 that intelligible electron-density maps, which are comparable to those obtained from 12 451 patterns by the Monte Carlo method in Figs. 6(*a*) and 6(*b*), are produced from only 1459 or 1452 patterns. The side-chain phases in the refined model are almost completely obtained after the G-correction step with the estimated crystal sizes obtained from either 2.5 or 5.0 Å resolution patterns. The differences between the electron-density maps calculated from these two situations in the improved integration method are barely visible. This indicates that an improved integration approach that requires fewer nanocrystals in SFX can be achieved by correcting for the geometric factor. Statistics of data simulation and structure refinement are summarized in Table 4.

To quantitatively assess the improvements in our integration method, we measured the standard linear correlation coefficient (CC) of the structure-factor amplitude and the average error in the phase angle ($\Delta\varphi$). Exact structure factors $F_{\rm exact}$ with phases $\varphi_{\rm exact}$ were calculated by the *CCP*4 *SFALL* program (Winn *et al.*, 2011) and recovered structure factors $F_{\rm calc}$ with phases $\varphi_{\rm calc}$ were measured by molecular replacement using the two integration methods from different numbers of images. CC and the $\Delta\varphi$ factor are defined as

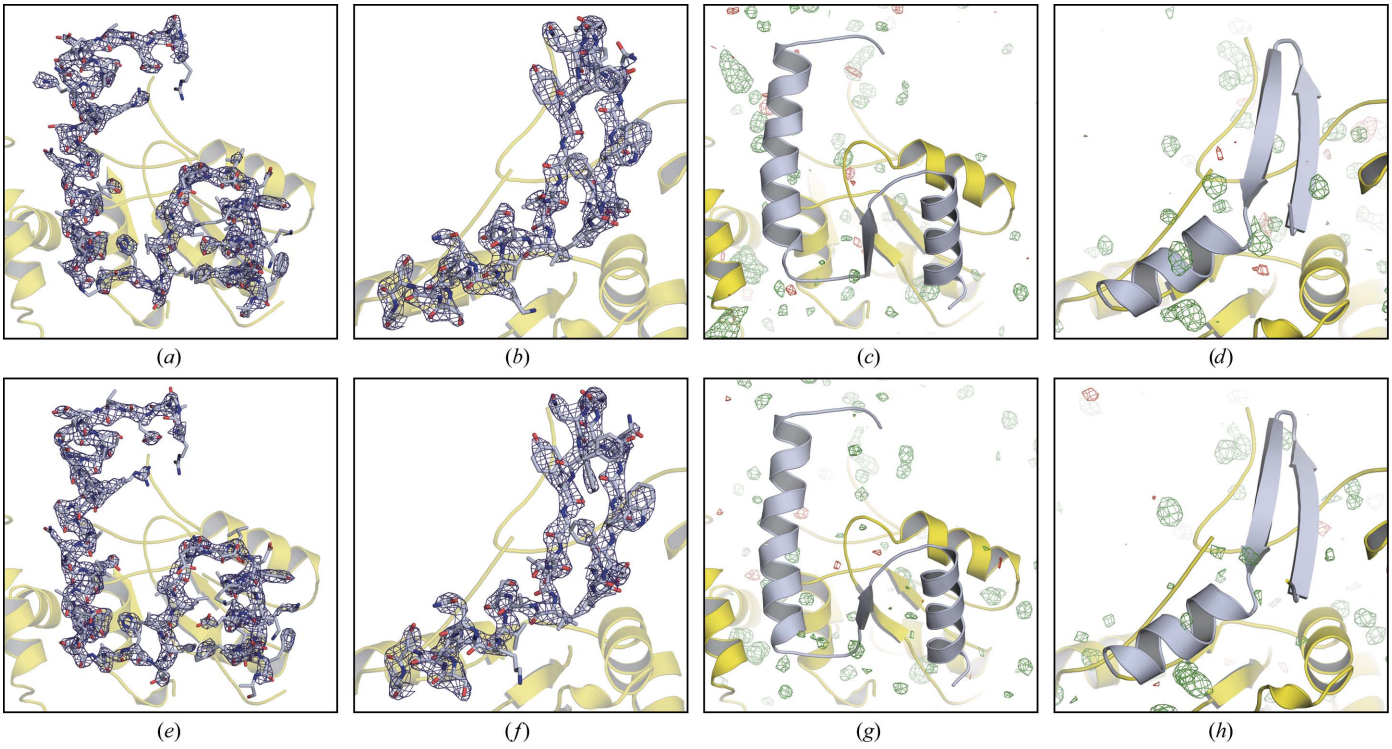


Figure 7
(*a*, *b*, *e*, *f*) Refined $F_{\rm obs} - F_{\rm calc}$ (blue, 2.5σ) electron-density maps of PDB entry 4f4j at 2.5 Å resolution by the improved integration method after the G-correction algorithm. (*c*, *d*, *g*, *h*) $F_{\rm obs} - F_{\rm calc}$ difference Fourier maps contoured at +3.5σ (green) and −3.5σ (red). (*a*, *b*, *c*, *d*) Correcting for the geometric factors by using the estimated crystal sizes reduced from 2.5 Å resolution patterns. 1459 corrected patterns were merged in the integration step. (*e*, *f*, *g*, *h*) Correcting for the geometric factors by using the estimated crystal sizes reduced from 5.0 Å resolution patterns. 1452 corrected patterns were merged in the integration step. Figures were produced using *PyMOL* (DeLano, 2002).

$$\mathrm{CC} = \frac{\sum_{hkl}(|F_{\mathrm{exact}}| - \langle|F_{\mathrm{exact}}|\rangle)(|F_{\mathrm{calc}}| - \langle|F_{\mathrm{calc}}|\rangle)}{\left[\sum_{hkl}(|F_{\mathrm{exact}}| - \langle|F_{\mathrm{exact}}|\rangle)^2 \sum_{hkl}(|F_{\mathrm{calc}}| - \langle|F_{\mathrm{calc}}|\rangle)^2\right]^{1/2}}, \tag{12}$$

and

$$\Delta\varphi = \frac{\sum_{hkl} \arccos|\cos(\varphi_{\mathrm{exact}} - \varphi_{\mathrm{calc}})|}{\sum_{hkl} 1}. \tag{13}$$

In Fig. 8, the resolution bin is evenly spaced into 20 shells. About 1260 structure factors fall into each shell. The horizontal axis represents the left end points of each resolution interval. The CC of the improved integration method is globally higher than that of the Monte Carlo integration method with 2000 patterns, and $\Delta\varphi$ drops significantly. This illustrates that the improved approach works successfully with the estimated crystal sizes and speeds up the convergence of CC and $\Delta\varphi$. However, because of the gap in the data completeness between the improved integration method and the Monte Carlo integration method with 12 451 patterns, the improvements are less remarkable in the high-resolution shells. This problem may be solved by adding hundreds of patterns, which is currently under investigation as a further study.

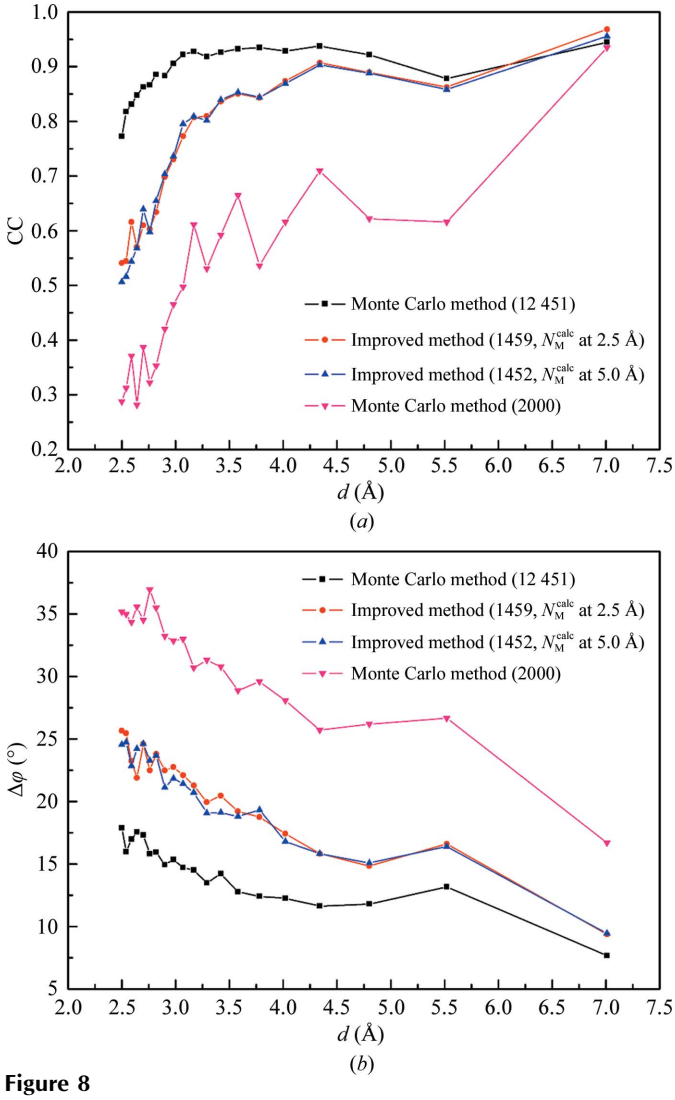


Figure 8
The standard linear correlation coefficient (*a*) and the average error in the phase angle (*b*) plotted against resolution. Statistics of the improved integration method are analyzed. Images were corrected with estimated crystal sizes calculated from high-resolution and low-resolution patterns. The Monte Carlo integration method inputting 12 451 and 2000 images omitting the G-correction step is shown as a control.

In the improved integration method, there are few differences in CC or $\Delta\varphi$. The G-correction algorithm works well with estimated crystal sizes obtained from either 2.5 or 5.0 Å resolution data. Although the R_{split} factor of the corrected images using the estimated crystal sizes at 2.5 Å resolution is higher, the R_{work} and R_{free} factors of the refined model are lower, as seen in Table 4. Mutually matching estimated crystal sizes with diffraction intensities at the same resolution seems to aid the geometric factor correction before the integration step.

4. Conclusions

The basic feature of the improved algorithm is to extract and normalize the effective intensities by correcting for the geometric factors in the directly measured intensities in each diffraction pattern before integrating them, which eliminates the influence of crystal size variation and improves the integral precision of single patterns. By analyzing peak profiles between adjacent pixels on the detector, crystal sizes can be estimated. The G-search algorithm is not very sensitive to crystal orientation errors. Using the geometric factor correction algorithm accelerates the convergence of the integration method and helps to decrease the number of individual snapshot diffraction patterns that are required in SFX.

It has been proved that the existing precision of the autoindexing algorithm are sufficient for the improved integration algorithm. Although the G-correction algorithm increases the second indexing accuracy, it is not very necessary. The previously determined orientation in the first indexing step may be used instead to ensure the indexing success rate in SFX.

Finally, when facing the low throughput in the present data-collection strategy, it is helpful for SFX pre-processing to determine estimated nanocrystal sizes with near-regular shapes. The improved algorithm presented is a promising method for data integration in SFX. Experimental verifications are in progress.

This work was supported by the National Basic Research Program of China (2009CB918600) and the National Natural Science Foundation of China (10979005).

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Boutet, S. *et al.* (2012). *Science*, **337**, 362–364.
Chapman, H. N. *et al.* (2011). *Nature (London)*, **470**, 73–77.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* D**66**, 12–21.
DeLano, W. L. (2002). *PyMOL*. http://www.pymol.org.
DePonte, D. P., Weierstall, U., Schmidt, K., Warner, J., Starodub, D., Spence, J. C. H. & Doak, R. B. (2008). *J. Phys. D Appl. Phys.* **41**, 195505.
Drenth, J. (2007). *Principles of Protein X-ray Crystallography*, 3rd ed. New York: Springer.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Johansson, L. C. *et al.* (2012). *Nature Methods*, **9**, 263–265.
Johnston, C. A., Whitney, D. S., Volkman, B. F., Doe, C. Q. & Prehoda, K. E. (2011). *Proc. Natl Acad. Sci. USA*, **108**, E973–E978.
Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916–924.
Kirian, R. A., Wang, X., Weierstall, U., Schmidt, K. E., Spence, J. C., Hunter, M., Fromme, P., White, T., Chapman, H. N. & Holton, J. (2010). *Opt. Express*, **18**, 5713–5723.
Koopmann, R. *et al.* (2012). *Nature Methods*, **9**, 259–262.
Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **26**.
Leslie, A. G. W. (1999). *Acta Cryst.* D**55**, 1696–1702.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Philipp, H. T., Hromalik, M., Tate, M., Koerner, L. & Gruner, S. M. (2011). *Nucl. Instrum. Methods Phys. Res. A*, **649**, 67–69.
Powell, H. R. (1999). *Acta Cryst.* D**55**, 1690–1695.
Rossmann, M. G. & van Beek, C. G. (1999). *Acta Cryst.* D**55**, 1631–1640.
Strüder, L. *et al.* (2010). *Nucl. Instrum. Methods Phys. Res. A*, **614**, 483–496.
White, T. A., Kirian, R. A., Martin, A. V., Aquila, A., Nass, K., Barty, A. & Chapman, H. N. (2012). *J. Appl. Cryst.* **45**, 335–341.
Winn, M. D. *et al.* (1994). *Acta Cryst.* D**67**, 235–242.
Zhou, L., Liu, P. & Dong, Y.-H. (2013). *Chin. Phys. C*, **37**, 028101.