Acta Crystallographica Section A
Foundations and Advances
ISSN 2053-2733

Received 17 April 2014
Accepted 23 July 2014

An alternative method for data analysis in serial femtosecond crystallography

Tao Zhang,[a]* Yang Li[b]* and Lijie Wu[c]

[a]CAS Key Laboratory of Soft Matter Physics, Institute of Physics, Chinese Academy of Sciences, No. 8, South 3rd Street Zhong Guan Cun, Beijing, 100190, People's Republic of China, [b]Program in Cellular and Molecular Medicine, Boston Children's Hospital, Harvard Medical School, 3 Blackfan Circle, Boston, MA 02115, USA, and [c]National Center for Protein Science Shanghai, State Key Laboratory of Molecular Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai, 200031, People's Republic of China. Correspondence e-mail: zhangtao@cryst.iphy.ac.cn, yang.li@childrens.harvard.edu

Serial femtosecond crystallography (SFX) [Chapman *et al.* (2011), *Nature*, **470**, 73–77], based on the X-ray free-electron laser, is a new and powerful tool for structure analysis at atomic resolution. This study proposes an extrapolation method for diffraction data analysis on the basis of diffraction intensity distribution in reciprocal space. Results show that this new method can restore SFX simulation data to structure factors that are more consistent with the structures used in simulation.

1. Introduction

The method of serial femtosecond crystallography (SFX) (Chapman *et al.*, 2011) has been successfully applied to protein crystallography with X-ray free-electron laser (FEL) sources. In SFX, diffraction spots appear in a large number of diffraction images for any particular reflection and each image can then be individually indexed. This unique characteristic allows data integration from different diffraction images. Currently Monte Carlo integration (Kirian *et al.*, 2010) is the generally adopted method to carry out this task.

However, Monte Carlo integration should not be the only valid method for SFX data analysis. The emergence of other methods would be beneficial to enhance the comprehension and development of SFX. This study proposes an extrapolation method based on the diffraction intensity distribution of imperfect crystals in reciprocal space. Simulation data were used to test this method and satisfactory results were obtained. The accuracy of reflection data processed by using this new method was markedly improved.

2. Extrapolation method

When one Bragg lattice point crosses the Ewald sphere, diffraction occurs in the direction of the intersection. The Ewald sphere can be regarded as an ideal spherical shell, because of high monochromaticity and the low divergence of the X-ray laser pulse. However, only the Fourier transform of a perfect crystal is an ideal lattice. The energy distribution in reciprocal space after Fourier transformation does not converge perfectly but rather disperses because of crystal defects and the finite number of cell units. This effect is more evident in SFX because the crystal used is considerably smaller than conventional crystals. For each reflection, the final energy distribution in reciprocal space can be considered as a Gaussian function around the corresponding lattice point. To simplify the case, all crystals are deemed identical, which indicates that the energy distributions of each crystal in reciprocal space are the same, as shown in equation (1). For one Bragg lattice point (h, k, l), the energy is assumed to be proportional to the square of the modulus of the structure factor F_{hkl} and exponentially decreases with increasing distance from the lattice point.

In equation (1)

$$E_{hkl}(r) = \left|F_{hkl}\right|^2 \exp\left(\frac{-\left|r - r_{hkl}\right|^2}{\sigma^2}\right), \qquad (1)$$

r is the coordinate of any point in the reciprocal space, r_{hkl} is the coordinate of Bragg lattice point (h, k, l), F_{hkl} is the structure factor of the reflection at this Bragg lattice point and σ is the standard deviation of the Gaussian distribution.

In SFX, the absence of the oscillation observed during data collection when using a traditional X-ray diffraction data collection method causes the recorded intensity of a reflection to be a static section of the overall intensity, as shown in Fig. 1. Such intensity is proportional to the energy distribution in the area. The majority of the diffraction intensity falls into a small region around the point r' on the Ewald sphere that is close to the Bragg lattice point, because the energy distribution at this point decays exponentially. We suppose that for reflection *hkl* in the ith diffraction pattern, this section locates at position r_i', such that the intensity of *hkl* is given by

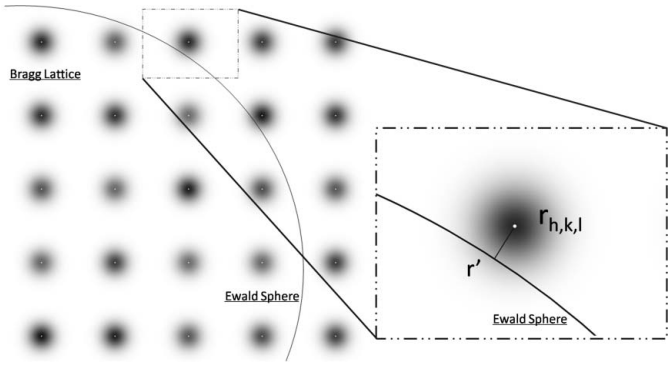

Figure 1
Ewald sphere and Bragg lattice in reciprocal space. The energy of a reflection spreads around the Bragg lattice point marked with a tiny white hollow. Diffraction occurs when the Ewald sphere is sufficiently close to a Bragg lattice point. The right-hand side is an enlarged view, where r' is a point on the Ewald sphere that has the shortest distance to a Bragg lattice point (h, k, l).

$$I_{hkl,i} \propto \left|F_{hkl}\right|^2 \exp\left(\frac{-\left|r'_i - r_{hkl}\right|^2}{\sigma^2}\right). \quad (2)$$

The scale of intensities in crystallographic structure analysis is arbitrary. Thus, equation (2) can be rewritten as

$$I_{hkl,i} = \left|F_{hkl}\right|^2 \exp\left(\frac{-\left|r'_i - r_{hkl}\right|^2}{\sigma^2}\right). \quad (3)$$

Each crystal is jetted into the beamline and then randomly hit by X-rays, such that the relative positions of the Ewald sphere and of the Bragg lattice point are randomly selected. The distribution of r' around one Bragg lattice point in reciprocal space is uniform and random. Based on these assumptions, a new extrapolation method to analyse SFX data is proposed and discussed.

The distribution of r' around one Bragg lattice point in reciprocal space is uniform and random, as discussed above. If the radii of the shells dividing the reciprocal space near a Bragg lattice point are an arithmetic progression (R, $2R$, ..., nR), r' will be uniformly distributed in these shells. The number of points in each shell is proportional to its volume. The volume of the first shell, which is a sphere, is given by

$$V_1 = \tfrac{4}{3}\pi R^3. \quad (4)$$

The volume of the second shell is the volume of a sphere with radius $2R$ from which the volume of the first shell is subtracted:

$$V_2 = \tfrac{4}{3}\pi(2R)^3 - \tfrac{4}{3}\pi R^3 = 7V_1. \quad (5)$$

Therefore, the volume of the nth shell is

$$V_n = \tfrac{4}{3}\pi(nR)^3 - \tfrac{4}{3}\pi[(n-1)R]^3 = (3n^2 - 3n + 1)V_1. \quad (6)$$

Accordingly, the intensities represented by the points in the inner shells are higher than those in the outer shells. In the proposed method, all intensities of a particular reflection are first sorted in ascending order. If m reflections with the highest intensities are located in the innermost shell, the next $7m$ intensities will be in the second shell and so on, such that all intensities are assigned to different shells. So the number of intensities m in the first shell is determined by the total quantity of measured intensities for one reflection, and the number of total shells n for a particular reflection is determined by m indirectly. Intensities in the same shell will evidently be close to one another. The mean value $I^n_{\rm mean}$ of all intensities in each shell is assumed to be the characteristic intensity of such a shell. The distance from the Bragg lattice point to the middle point of this shell, $R^n_{\rm c} = (n - 0.5)R$, is the characteristic radius of the shell. A characteristic couple $(I^n_{\rm mean}, R^n_{\rm c})$ can be obtained for the nth shell. One reflection will have a series of couples for different n. Taking the logarithm of both sides of equation (3) and substituting the characteristic couples in $(I^n_{\rm mean})$ to the intensity on the left and $R^n_{\rm c}$ to the distance from the Bragg lattice point (h, k, l), a series of equations can be obtained:

$$\ln\left(I^n_{\rm mean}\right) = \ln\left(\left|F_{hkl}\right|^2\right) - R^{n^2}_{\rm c}/\sigma^2. \quad (7)$$

In these equations, $|F_{hkl}|$ and σ are constants. The least-squares method was applied to minimize the target function equation (8) and to calculate the best estimates of $\ln(|F_{hkl}|^2)$ and $1/\sigma^2$. Consequently, the square of the modulus $|F_{hkl}|^2$ can be obtained by using this method. Thus, equation (7) was extrapolated to $R^n_{\rm c} = 0$:

$$\Psi = \sum\left\{\ln\left(I^n_{\rm mean}\right) - \left[\ln\left(\left|F_{hkl}\right|^2\right) - R^{n^2}_{\rm c}/\sigma^2\right]\right\}^2. \quad (8)$$

In practice, the squares of the radii form an arithmetic progression. The radius of the nth shell is set to $n^{1/2}R$ instead of nR. Thus, the volume of the first shell remains the same, whereas that of the nth shell is changed to

$$\begin{aligned} V_n &= \tfrac{4}{3}\pi\left(n^{1/2}R\right)^3 - \tfrac{4}{3}\pi[(n-1)^{1/2}R]^3 \\ &= [n(n)^{1/2} - (n-1)(n-1)^{1/2}]V_1. \end{aligned} \quad (9)$$

The number of intensities in each shell must be modified accordingly. If m intensities exist in first shell, the second shell should contain $[2(2)^{1/2} - (2-1)(2-1)^{1/2}]m = 1.83m$ intensities, and the third shell should contain $[3(3)^{1/2} - (3-1)(3-1)^{1/2}]m = 2.35m$. The characteristic value of R^2 for the nth shell is $(R^2)^n_{\rm c} = (n - 0.5)R^2$ and the target function is given by

$$\Psi = \sum\left\{\ln\left(I^n_{\rm mean}\right) - \left[\ln\left(\left|F_{hkl}\right|^2\right) - (R^2)^n_{\rm c}/\sigma^2\right]\right\}^2. \quad (10)$$

Compared with $R^{n^2}_{\rm c} = (n - 0.5)^2R^2$ in equation (8), $(R^2)^n_{\rm c}$ has equidistant intervals for equation (10), which can improve the efficiency of the least-squares method.

A series of points representing couples $\left[\ln\left(I^n_{\rm mean}\right), \left(R^2\right)^n_{\rm c}\right]$ for four reflections from the *E. coli* DhaR(N)–DhaL complex are shown in Fig. 2. The straight line is based on equation (11), which was derived from equation (1). Parameters $\ln\left(|F_{hkl}|^2\right)$ and $1/\sigma^2$ were determined by minimizing equation (10). Therefore, $\ln(I)$ is an expected value on the basis of the intensity distribution and is given by

$$\ln(I) = \ln\left(\left|F_{hkl}\right|^2\right) - \left(R^2\right)_c/\sigma^2. \qquad (11)$$

The spots represent couples $\left[\ln\left(I^n_{\rm mean}\right), \left(R^2\right)^n_c\right]$ and $\ln\left(I^n_{\rm mean}\right)$ is calculated from experimental measurements. Regardless of the chosen $\left(R^2\right)_c$ value, experimental values are all close to the expected value, which proves that the experimental intensity fall-off of a reflection in reciprocal space coincides with equation (1).

Table 1
Summary of crystallographic information for the samples used in the simulation.

AU: asymmetric unit.

	Apo GroEL	*E. coli* DhaR(N)–DhaL complex	Lysozyme
Residues per AU	7658	2116	129
Unit-cell parameters			
a, b, c (Å)	262.80, 283.60, 135.72	89.77, 91.50, 93.81	72.29, 72.29, 38.12
α, β, γ (°)	90.0, 90.0, 90.0	84.15, 72.42, 90.01	90.0, 90.0, 90.0
Space group	$P2_12_12_1$	$P1$	$P4_32_12$
PDB code	1xck	4lrz	4etc

3. Samples

Simulation data were used to test the proposed method. Such data can be used to determine the true intensity values which would be known exactly in a simulation, which is an absolute standard for judging the quality of processed reflection data, hence the efficiency of the analysis method. Meanwhile, real experimental data have numerous error sources that impede analysis and judgement.

Three proteins were used in the simulation. Apo GroEL (Bartolucci *et al.*, 2005) is a huge protein with a large unit cell. The *E. coli* DhaR(N)–DhaL (Shi *et al.*, 2014) complex has a low crystallographic symmetry, with space group $P1$. The structure of lysozyme was determined on the basis of room-temperature data collected at the Swiss Light Source (SLS), which are a reference for high-resolution protein structure determination by serial femtosecond crystallography (Boutet *et al.*, 2012). The details of these three proteins are listed in Table 1. The coordinate files of the samples were downloaded from the Protein Data Bank (Berman *et al.*, 2000).

4. Process

4.1. Simulations

For Apo GroEL, the simulation parameters followed *lcls-xpp-estimate.beam* and *lcls-xpp-estimate.geom*, which were released with the *CrystFEL* package (White *et al.*, 2012). We simulated an XFEL pulse with wavelength λ = 4.13 Å (3000 eV) and 1×10^{12} photons per pulse. The detector

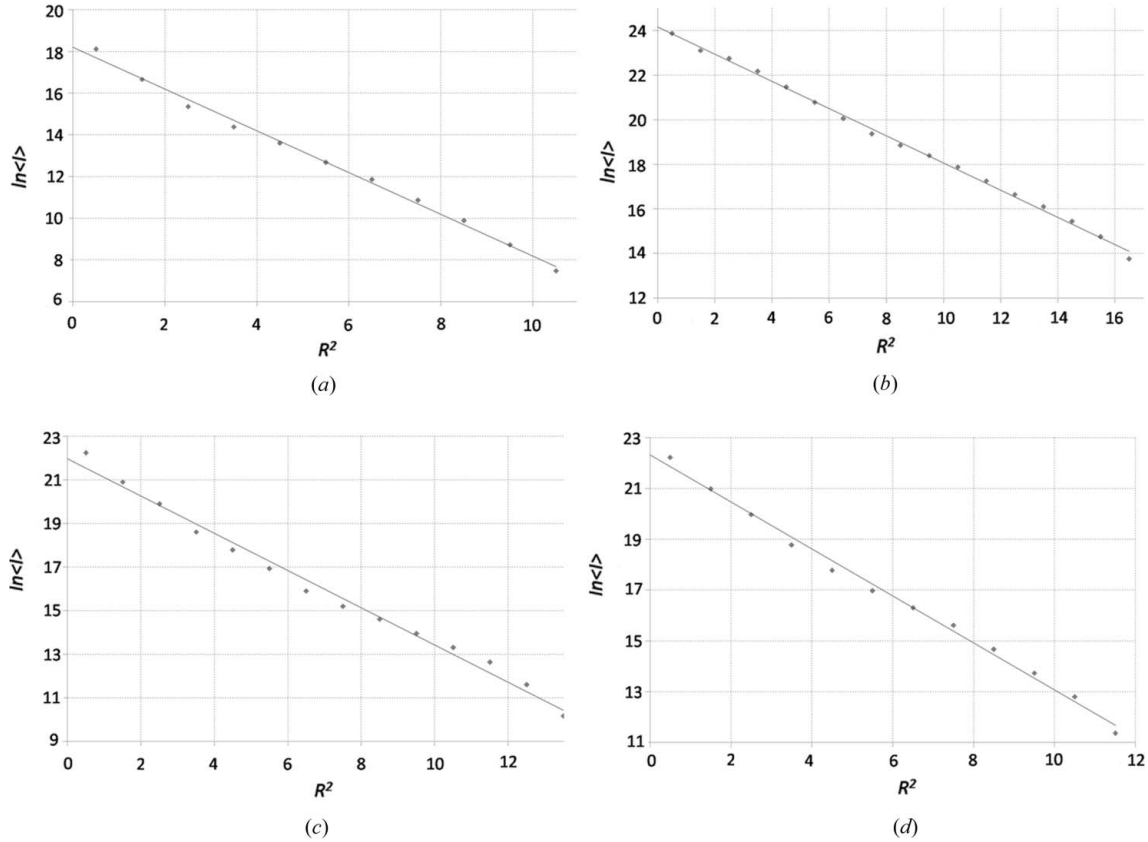

Figure 2
Characteristic couples $\left[\ln\left(I^n_{\rm mean}\right), R^{n^2}_c\right]$ and the straight line based on equation (11) for four reflections (*a*) $9\bar{1}0$, (*b*) $13,\bar{9},1$, (*c*) 2,6,30 and (*d*) 16,18,2. The data are from the simulation data of the *E. coli* DhaR(N)–DhaL complex. The slope $1/\sigma^2$ and intercept $\ln(|F_{hkl}|^2)$ of the straight line were the best estimates of the least-squares method.

contained an array of 1456 × 1456 pixels of individual length 110 µm and the sample-to-detector distance was set at 8 cm. To obtain high-resolution diffraction data while considering small unit cells, the wavelength was changed to λ = 1 Å (12400 eV) and the sample-to-distance was reduced to 2 cm for the *E. coli* DhaR(N)–DhaL complex and lysozyme.

Only one crystal was hit once by the X-ray laser. The parameters of the X-ray laser for each hit remained unchanged, while the sizes of crystals differed from 500 nm to 5000 nm. This range corresponds to 20 to 200 unit cells in each dimension for Apo GroEL and 60 to 600 for the other two samples, which is similar to the real experiment. Poisson noise was added during the simulation. The simulations were performed with the use of *pattern_sim*, a program of *CrystFEL*.

4.2. Data analysis

Diffraction patterns were indexed one by one by using *indexamajig* in *CrystFEL*. In this program, diffraction peak positions were determined by using the built-in Zaefferer algorithm and passed on to *DIRAX* (Duisenberg, 1992) and *MosFLM* (Collaborative Computational Project, Number 4, 1994) for indexing. The intensities of these indexed diffraction peaks were normalized by using a two-pass process. The intensities were scaled to the most consistent values produced by a previous Monte Carlo integration with unscaled patterns (White *et al.*, 2012). The Monte Carlo integration is used only for data scaling without the need to generate any data for structure analysis. The proposed method was then used to integrate the scaled intensities and to generate the final reflection data. For comparison, Monte Carlo integration was also applied to generate a separate data set, which was implemented by *process_hkl* in the *CrystFEL* package. During extrapolation, intensities with zero or negative values were discarded because the logarithm of intensities was used in the calculation. To facilitate the comparison, two data sets were generated by using Monte Carlo integration. In one data set, zero and negative intensities were retained, whereas zero and negative intensity measurements were omitted from the other data set. The program *SOLVE* in the *PHENIX* package (Adams *et al.*, 2010) was employed to calculate the structure factors from intensities.

5. Result

The *R* factor calculated with the structure used in simulation and calculated structure factors was adopted as an indicator for the two different methods because the structure used in simulation is an absolute standard for assessing the quality of reflection data. *REFMAC5* (Murshudov *et al.*, 2011) was used to calculate the *R* factor without any refinement.

The line chart of *R* factor *vs* resolution is shown in Fig. 3. In this calculation, 1×10^5 diffraction patterns were used for Apo GroEL and lysozyme, while tripled patterns (3×10^5) were applied to *E. coli* DhaR(N)–DhaL complex to compensate for the low symmetry. The results of the new method are better than those of Monte Carlo integration, except for only two points in the low-resolution region of lysozyme. The gap is very clear at most resolution ranges. Although the results of two kinds of Monte Carlo integration (with and without zero or negative intensities) do not exhibit noticeable differences, the *R* factor of Monte Carlo integration using only positive intensities is slightly higher than that of Monte Carlo integration with all intensities. The number of reflections used to calculate every *R* factor is represented in the histogram. The overall *R* factors are listed in Table 2. The overall *R* factors of the new method are significantly lower than those of Monte

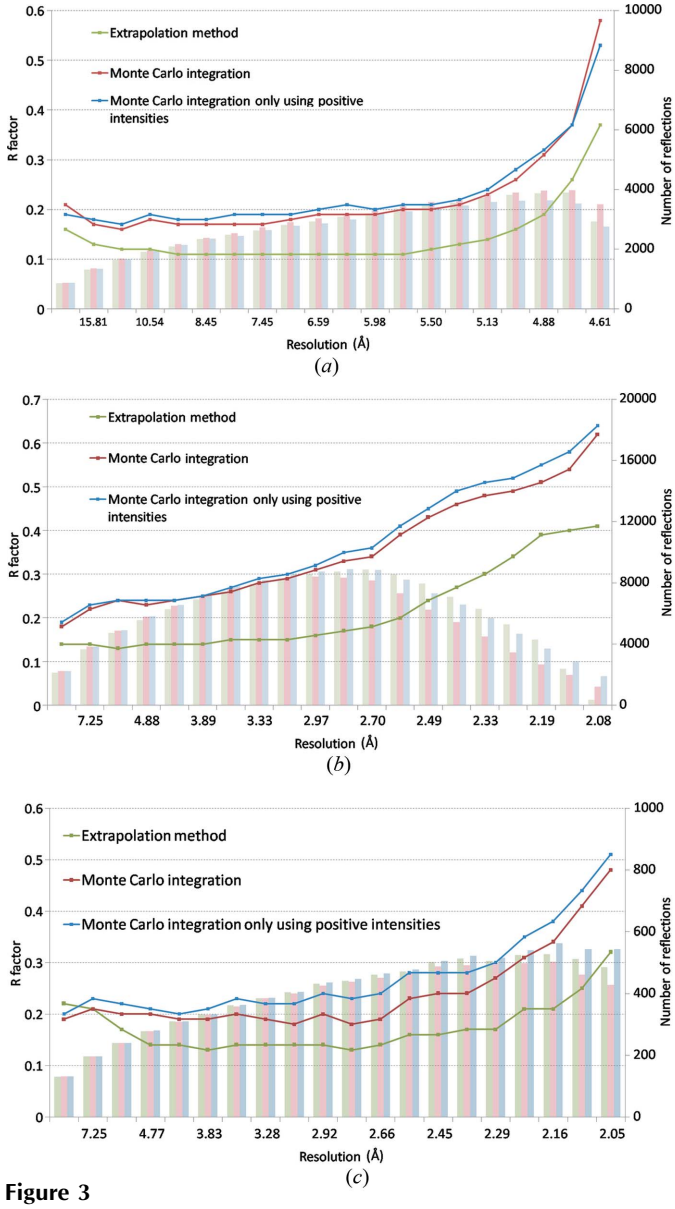

Figure 3
R factor *vs* resolution. The green line represents the data processed by using the extrapolation method. The blue and red lines represent Monte Carlo integration with and without zero or negative intensities omitted respectively. The histograms represent the number of reflections used to calculate the *R* factor. (*a*) Apo GroEL, (*b*) *E. coli* DhaR(N)–DhaL complex and (*c*) lysozyme.

Carlo integration. These results show that the reflection data processed by the new method are closer to the original structure.

The percentages of intensities excluded because of non-positivity for each of the samples are shown in Fig. 4. The high resolution results in a large percentage of intensities that have zero and negative values. These intensities are excluded in both the extrapolation method and Monte Carlo integration to facilitate comparison. For the *E. coli* DhaR(N)–DhaL complex and lysozyme, over half of the intensities were discarded when the resolution was close to 2.0 Å.

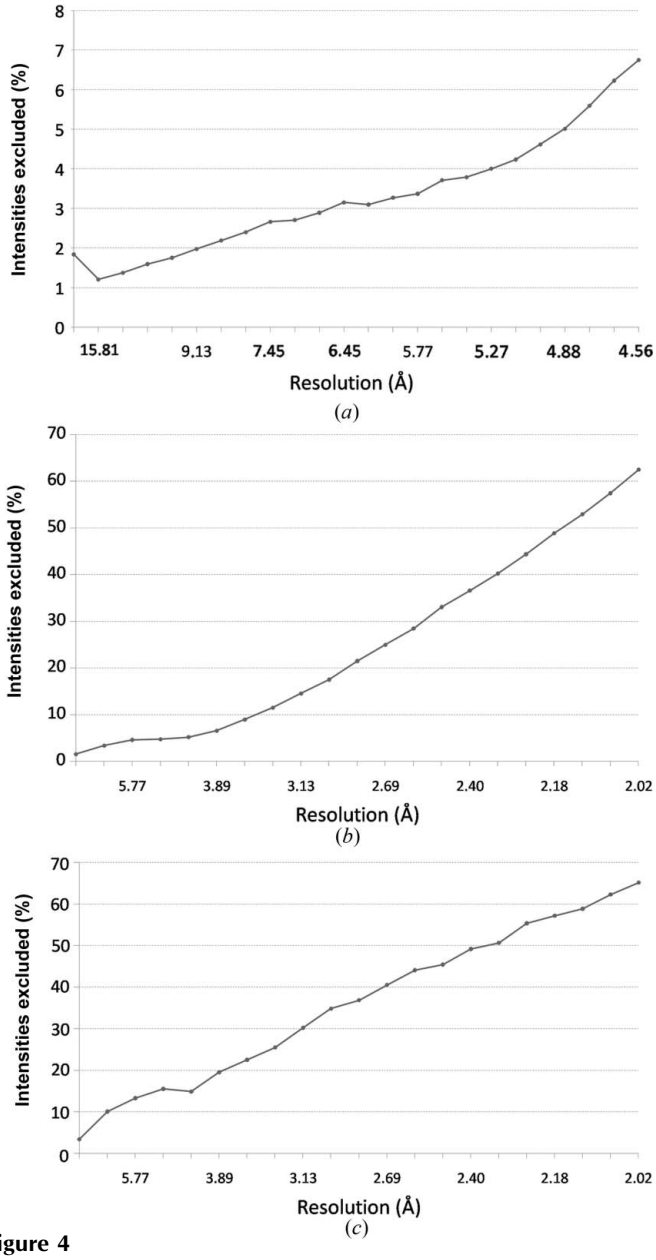

Figure 4
Percentage of intensities with zero or negative values *vs* resolution, which would be discarded in the extrapolation method and Monte Carlo integration to facilitate comparison. (*a*) Apo GroEL, (*b*) *E. coli* DhaR(N)–DhaL complex and (*c*) lysozyme.

Table 2
Overall *R* factors.

	Apo GroEL	*E. coli* DhaR(N)–DhaL complex	Lysozyme
Extrapolation method	0.1422	0.1649	0.1712
Monte Carlo integration	0.2210	0.2917	0.2266
Monte Carlo integration using only positive values	0.2294	0.3142	0.2654

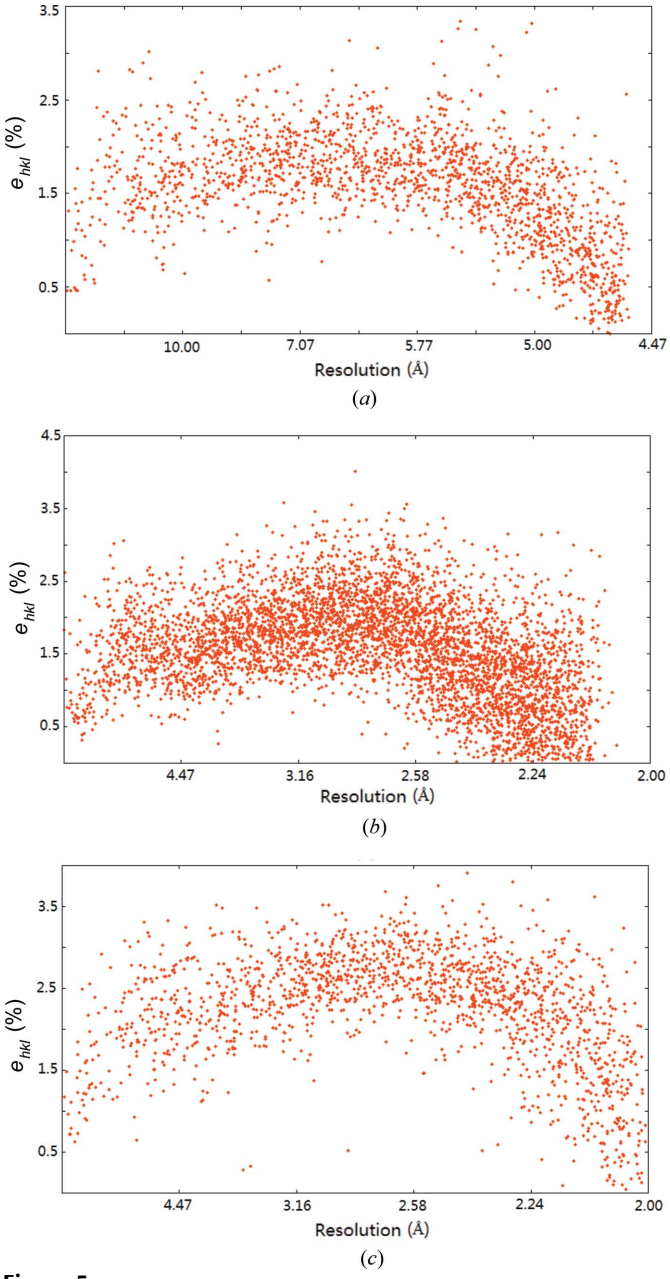

Figure 5
Plot of e_{hkl} in equation (12) *vs* resolution, where e_{hkl} is an indicator used to measure agreement between real conditions and the Gaussian intensity distribution. To demonstrate the effect better, only some of the reflections (selected at random) are shown in the figure. (*a*) Apo GroEL (2% of reflections were selected), (*b*) *E. coli* DhaR(N)–DhaL complex (5%) and (*c*) lysozyme (20%).

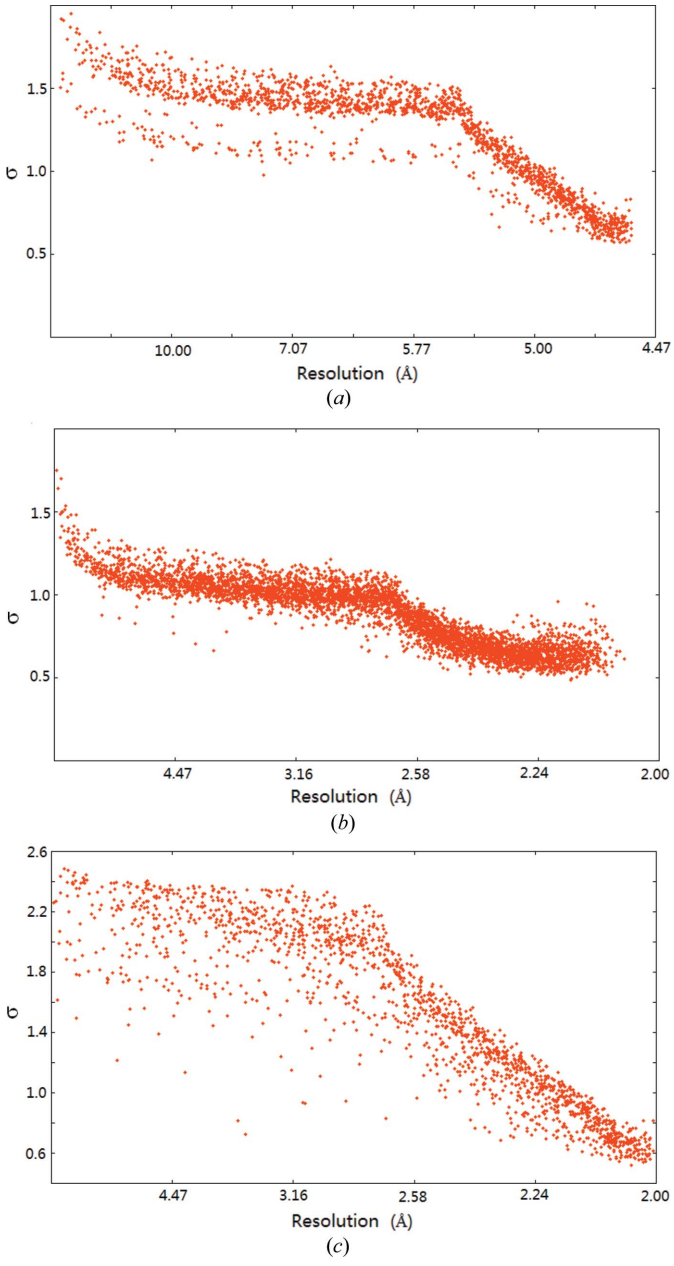

Figure 6
Value of σ in equation (7) for each reflection *vs* resolution, where σ is the standard deviation of the Gaussian intensity distribution. To demonstrate the effect better, only some of the reflections (selected at random) are shown in the figure. (*a*) Apo GroEL (2% of reflections were selected), (*b*) *E. coli* DhaR(N)–DhaL complex (5%) and (*c*) lysozyme (20%).

In equation (12),

$$e_{hkl} = \frac{\left\langle \left| \ln\left(I^{n}_{\rm mean}\right) - \left[\ln\left(\left|F_{hkl}\right|^{2}\right) - \left(R^{2}\right)^{n}_{\rm c}/\sigma^{2} \right] \right| \right\rangle}{\ln\left(\left|F_{hkl}\right|^{2}\right)}, \qquad (12)$$

e_{hkl} is used as an indicator to measure agreement between the real condition and the proposed intensity Gaussian distribution. Reflections plotted against resolutions are shown in Fig. 5 for the three samples. Almost all values of e_{hkl} are less than 3.5% for any sample, which indicates that the real intensity distribution effectively fits the Gaussian distribution in reciprocal space.

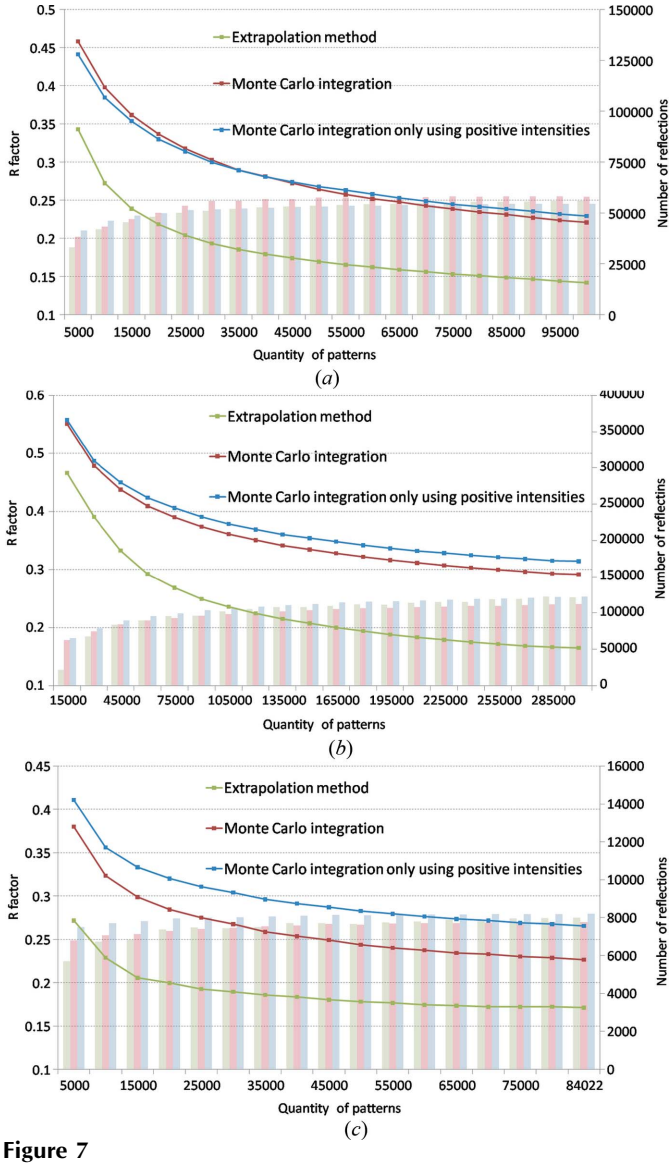

Figure 7
Overall *R* factor *vs* quantity of diffraction patterns used. The green line represents the data processed by using the extrapolation method. The blue and the red lines represent Monte Carlo integration with and without zero or negative intensities, respectively. The histograms represent the number of reflections used to calculate the *R* factor. (*a*) Apo GroEL, (*b*) *E. coli* DhaR(N)–DhaL complex and (*c*) lysozyme.

The value of the deviation in equation (7) for each reflection *vs* resolution is shown in Fig. 6. The values with the same resolution are centralized in a small range, especially for the *E. coli* DhaR(N)–DhaL complex, which may be affected by various mechanisms and will be the focus of future research.

The relationship between the quantity of diffraction patterns and quality of reflection data was also investigated for different methods. Fig. 7 shows the line charts of the relationships. For any number of patterns, *R* factors from the new method are lower than those from the two types of Monte Carlo integration. As expected, more diffraction patterns

improved the results of any method. The numbers of reflections are also shown in the histograms.

6. Conclusion

In this study, an extrapolation method for SFX data analysis was proposed by analysing the diffraction intensity distribution in reciprocal space. The principle of this new method is different from that of Monte Carlo integration. The proposed method generates accurate reflection data according to simulation results. This method is a useful option for SFX analysis and has the potential to produce satisfactory results for real experimental data.

This work was supported by the National Natural Science Foundation of China (Grant No. 11204364). We thank Professor Haifu Fan for discussions.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Bartolucci, C., Lamba, D., Grazulis, S., Manakova, E. & Heumann, H. (2005). *J. Mol. Biol.* **354**, 940–951.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Boutet, S. *et al.* (2012). *Science*, **337**, 362–364.
Chapman, H. N. *et al.* (2011). *Nature*, **470**, 73–77.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Duisenberg, A. J. M. (1992). *J. Appl. Cryst.* **25**, 92–96.
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.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Shi, R., McDonald, L., Cygler, M. & Ekiel, I. (2014). *Structure*, **22**, 478–487.
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.