New data reduction protocol for Bragg reflections observed by TOF single-crystal neutron diffractometry for protein crystals with large unit cells

Katsuaki Tomoyori and Taro Tamada
Quantum Beam Science Directorate, National Institutes for Quantum and Radiological Science and Technology
2-4 Shirakata, Tokai, Naka, Ibaraki 319-1106, Japan

E-mail: tomoyori.katsuaki@qst.go.jp

Abstract. In protein crystallography, high backgrounds are caused by incoherent scattering from the hydrogen atoms of protein molecules and hydration water. In addition, the scattering intensity from large unit-cell crystals is very small, which makes it difficult to improve the signal-to-noise ratio. In the case of time-of-flight (TOF) single-crystal neutron diffractometry, the measured spectra cover four-dimensional space including X, Y, and TOF in addition to intensity. When estimating the integrated intensity, 3D background domains in the vicinity of peaks should be clearly classified. In conventional 1D or 2D background evaluation, the evaluation is applied for individual peaks assigned using peak searches; however, it is quite difficult to classify the 3D background domain in TOF protein single-crystal neutron diffraction experiments. We undertook the development of a data reduction protocol for measurements involving large biomacromolecules. At the initial stage of the reduction protocol, appropriate 3D background estimation and eliminations were applied over the entire range of X, Y, and TOF bins. The histograms were then searched for peaks and indexed, and the individually assigned peaks were finally integrated with an effective profile function in the TOF direction. Three-dimensional deconvolution procedures for overlapping peaks associated with large unit cells were implemented as necessary. This data reduction protocol may lead to the improvement of signal-to-noise ratios to enable TOF spectral analysis of large unit-cell protein crystals.

1. Introduction

Neutron protein crystallography has the unique ability to locate the precise positions of hydrogen atoms and hydration water molecules in biomacromolecules. We will construct a high-resolution time-of-flight (TOF) protein single-crystal neutron diffractometer at the Materials and Life Science Experimental Facility (MLF) at the Japan Proton Accelerator Research Complex (J-PARC). Using a spallation neutron source, the diffractometer can efficiently sweep a large reciprocal space in a single measurement via a white thermal neutron beam, which is suitable for diffraction experiments. This new diffractometer will be designed to be the best-in-class for efficient measurement of Bragg data sets at a resolution of 2.0 Å for protein crystals with unit cell constants of 250 Å. These include over 95% of the protein structures, including large proteins such as membrane proteins and protein complexes, deposited in the Protein Data Bank [1]. We will achieve high throughput, minimal peak overlap, and high signal-to-noise ratio by using a large wavelength bandwidth of TOF-sorted neutrons and an array of high-spatial-resolution position-sensitive area detectors covering a large solid angle. Large unit-cell protein crystals lead to weak average peak intensities. In addition, high backgrounds

are caused by incoherent scattering from the hydrogen atoms of protein and hydration water molecules. To precisely calculate weak signals in the presence of such high backgrounds, a data reduction protocol is needed to improve the signal-to-noise ratio and thus enable the neutron structural analysis of large proteins.

Two types of procedures exist for the integration of Bragg reflections: the box summation method, which simply designates each peak region within a box, and the profile fitting method. Box summation is frequently used to integrate a designated domain while distinguishing between signals and background. Having determined the background domain, a simple estimate of the integrated intensity is obtained by summing the pixel values of all the pixels in the peak domain of the mask and then subtracting the sum of the background values calculated from the background domain for the same pixels. If the background level is very low compared with the intensity of a spot and the spots are well resolved, this will give as accurate an estimate of the intensity as possible.

In TOF neutron protein crystallography, the intensities of the diffraction spots can be integrated using STARGazer [2], which employs box summation strategies, at MLF/J-PARC, using subroutines from the Integrated Spectral Analysis Workbench (ISAW) software [3]. To carry out the integration in TOF single-crystal diffractometry involving four-dimensional (4D) spaces with XY + TOF plus intensity, the 3D classified domains enclosed by XY + TOF around each spot are box-summed and computed after background elimination. It is important that pixels are not misclassified, as misclassification can lead to systematic errors in the integrated intensity. There are weak Bragg signals with asymmetric shapes over a large, structured, incoherent backgrounds derived from the incident profile in the TOF direction. In addition, specific backgrounds associated with the X-Y accidental coincidence signal are non-negligible. The large variety of continuous background shapes in the vicinity of the peaks should be faithfully reproduced over the entire range of space and time directions for every peak. However, conventional background estimation is inadequate for improving the signal-to-noise ratios of peak integration, especially for weak reflections.

To achieve high signal-to-noise ratios for TOF protein single-crystal neutron diffractometry with large unit-cell proteins, we investigated the methodology of background evaluation and ways of implementing the background algorithm in the data reduction protocol. The optimization of the data reduction protocol, combined with the use of the background estimation and the profile-fitting and deconvolution methods, may lead to the improvement of signal-to-noise ratios to enable TOF spectral analysis of large unit-cell protein crystals.

2. Data reduction protocol targeted for protein crystals with large unit cells

In protein crystallography, there are two distinct procedures for integrating spot intensities, known as 2D (XY) and 3D (XY + "Frame") integration, available in different software packages; e.g., MOSFLM [4] and HKL2000 [5] use the 2D (XY) method, while d*TREK [6] and XDS [7] use the 3D method. The 3D (XY + "Frame") indicates that the data set consists of multiple 2D (XY) data sets in which Bragg spots are recorded. The spectra themselves are 2D (XY) images including only 2D spatial coordinates.

Fig. 1 shows the conventional reduction protocol for protein crystallography. Using the orientation matrix, UB, we can predict where single crystal peaks should be found in detector/TOF space. The UB describes the sample orientation with respect to the diffractometer angles. Given UB, it is possible to calculate the diffractometer angles necessary to rotate a particular scattering vector **Q**, indexed by (*h, k, l*), into the diffraction position. The matrix **B** transforms the given (*h, k, l*) into an orthonormal coordinate system fixed in the crystal. The matrix **U** is the rotation matrix that rotates the crystal's reference frame into the spectrometer's [8]. Providing the background and peak regions are correctly defined, summation integration provides a method for evaluating integrated intensities which is both robust and free from systematic error. For weak reflections, however, many of the pixels in the peak region will contain very little signal (Bragg intensity), but will contributes significantly to the noise because of the Poissonian variation in the background. It is possible to obtain a more accurate estimate

of the integrated intensity by using a procedure known as profile fitting. In this method, it is assumed that the shape or profile (in 2D/3D) of the spots is known. The background plane is determined in the same way as for box summation, but the intensity is derived by determining the scale factor, which gives the best fit to the observed spot profile when applied to the known spot profile. This scale factor is then proportional to the profile-fitted intensity for the reflection. However, the 4D (XY + TOF plus intensity) precise summation integration or profile-fitting approach has not yet been used to carry out data reduction of TOF protein single-crystal neutron diffraction data. Herein, the notation "4D" indicates four-dimensional spaces including X, Y, and TOF in addition to intensity.

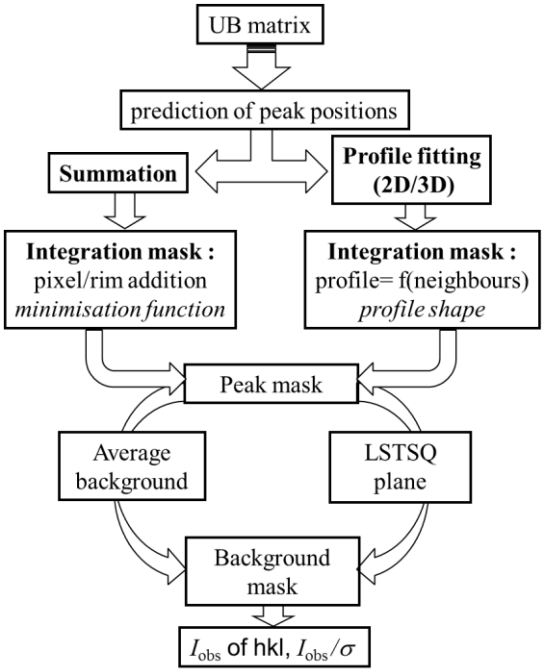

Fig 1. Conventional reduction protocol for protein crystallography

To apply the 4D Laue neutron diffraction data to the conventional data reduction protocol, as shown Fig. 1, a suitable choice of 3D background domain or 3D signal shape functions around spots is required to carry out precise peak integration of the Bragg reflections. In spot-based background estimation, the background around the peak spots is locally averaged or calculated by least-square method (LSTSQ) inside a designated 3D domain. In protein crystallography, incoherent neutron scattering events due to the hydrogen atoms of protein and surrounding water molecules are continuously distributed over the entire range of the measured TOF region. On the incoherent backgrounds, there are coherent weak Bragg signals, and the peaks exhibit asymmetric tails in the TOF direction due to the neutron moderation process used to extract the available thermal neutrons for diffraction experiments. The tails also prevented us from accurately discriminating between signal and background events. Furthermore, we will employ the 2D neutron detectors in X-Y coincidence techniques. Therefore, the accidental coincidence events associated with signal processing are continuously present as the backgrounds of the 3D Bragg peaks in the spatial directions. Therefore, the conventional method for designating the 3D background domain in the vicinity of the peaks cannot accurately describe the large variety of continuous background shapes underlying a single peak in both spatial and time directions, indicating that conventional background estimation is quite inadequate for improving the signal-to-noise ratio of peak integration for weak reflections in particular. It is clearly important that pixels are not misclassified, as this can lead to systematic errors in the integrated intensity. It is essential to employ a highly precise background evaluation method that can cope with the backgrounds of weak reflections. One or two dimensions can be reduced from the original dimensions to evaluate backgrounds. However, projecting the histogram data onto the lower dimension leads to degradation of the information between signal and background, resulting in systematic errors in the integrated intensity.

We will intend to decrease the misclassification by appropriately implementing the background elimination procedures with the intact 4D spectra, which apart from peak integration procedure with the profile functions to reproduce the asymmetric tail of Bragg reflections. The optimization of the data reduction protocol tailored to cope with 4D spectra observed by Laue single crystal neutron diffractometry could be inevitable to improve the signal-to-noise ratio. Profile fitting provides a means of improving the signal-to-noise ratio especially for large unit-cell protein crystals. The integration of weak reflections by using profile fitting has been demonstrated. In TOF protein single-crystal neutron diffratometry, we speculate an improvement of totally about 10% signal-to-noise ratio can be expected: a few percent for the discrimination on asymmetric tail, and a few percent for the

discrimination on spatial and time direction, respectively. In addition, the approaches based on the scaling of learned peak shapes found from strong neighboring peaks as shown Fig. 1 further could improve the signal-to-noise ratio.

When dealing with continuously varying backgrounds, non-iterative background approximation methods such as the three-window method [9] or polynomial fitting [10] are used. In addition, iterative methods such as simple multi-point smoothing [11] or more sophisticated methods such as the statistics-sensitive non-linear iterative peak-clipping (SNIP) algorithm [12] are used. The SNIP algorithm can iteratively compute optimum continuously varying backgrounds in accordance with the peak widths and is a good method for multidimensional spectra [13]. This algorithm can compute 3D background domains without decreasing dimensions, that is, losing background information. Recently, we applied it to the observed Bragg reflections measured by TOF protein single-crystal neutron diffractometry and found it to be an appropriate method for realizing the criteria mentioned above [14].

Fig. 2 shows a data-flow diagram of the data reduction protocol used in the TOF protein single-crystal neutron diffraction experiment. The left figure shows a typical reduction protocol and the right figure shows the proposed new reduction protocol, which is suitable for the integration of diffraction sets from large-unit-cell proteins. In a typical reduction protocol, the event data from a sample, collected by detectors, were presented as a histogram. Vanadium spectra measured under identical conditions were used to normalize the sample spectra. Vanadium is a purely incoherent scatter and incoherent scattering is by definition isotropic. The scattering intensity seen in each detector is a measurement of detector efficiency, solid angle and analyzer transmission. Vanadium scattering ordinarily takes into account the relative efficiencies of different pixels on the detector, the incident neutron flux profile, I_0 (λ), and the detector response at different neutron wavelengths (I_0 (λ) and detector sensitivity correction in Fig. 2). Each 3D data histogram was searched for peaks and these were indexed using a UB matrix. A refined UB matrix was used to determine which (X, Y, TOF) points should be included in the (q_h, q_k, q_l) points in the reciprocal lattice plane of interest. With many reflections determined within each frame, this allowed straightforward determination and refinement of the UB matrix. The cell dimensions were determined using a refinement of the UB matrix with respect to a set of reflections taken from a wide range of data frames, thus ensuring good averaging. For peak integration, on the other hand, a local UB refined for the particular frame under consideration was used. After refinement of the local UB, peak integration was undertaken in one of two ways, using either the peak position predicted from the UB matrix or those found during peak searching. After the peak position was predicted, the peak domain was determined and the background domain was also designated to evaluate the local background in the vicinity of a peak. The backgrounds were subtracted and the peak was box-summed or projected and integrated in a 1D TOF direction with appropriate asymmetric functions [15]. The integrated intensity was finally estimated for each Bragg reflection. The resulting intensities were reduced to structure factors, giving data sets of reflections after various basic corrections associated with the measurements and the principle.

In the proposed new reduction protocol, after correcting the incident neutron flux profile and the pixel and wavelength efficiency of the detectors, as mentioned in the typical reduction protocol, the entire 3D background estimation and subtraction were applied at the initial stage of the data reduction processes (the green colored component in Fig. 2, right), although background estimation was carried out for individual spot areas predicted by the refined UB matrix in a typical protocol. After the background elimination and peak identification processes were complete, the peaks found were then indexed and evaluated either as legitimate peaks or as overlapping peaks of higher order reflections. The potential overlapping peaks were routed through a deconvolution process, which is an efficient algorithm for multidimensional deconvolution (i.e. the Gold algorithm contained in TSpectrum [16]), and re-evaluated (the blue colored component in Fig. 2, right). Each of the single peak regions convolved in a doublet was resolved within an appropriate accuracy after iteration of the Gold deconvolution. The indices were then assigned for the resolved single peaks (Fig. 1 (1)). The fitting parameters including analytical functions thus employed were reflected on the integration processes (Fig. 1 (2)). No background mask was used when integrating the individual peaks (the yellow

component in Fig. 2, right); instead, background estimation and elimination were applied at the initial stage. The neutron peak counts for each spot could be summed in the appropriate peak mask domain. Otherwise, the spot could be projected onto a 1D/2D histogram and integrated using a profile fitting approach based on the known analytical shape of the reflections in the TOF direction, which were well understood from the characteristics of the neutron source. We found that the Landau–Vavilov distributions, which are used to describe the energy loss of charged particles traversing a thin absorber, were in excellent agreement with the observed TOF profile measured using an IBARAKI Biological Crystal Diffractometer (iBIX) installed at a coupled moderator source at MLF/J-PARC [17].

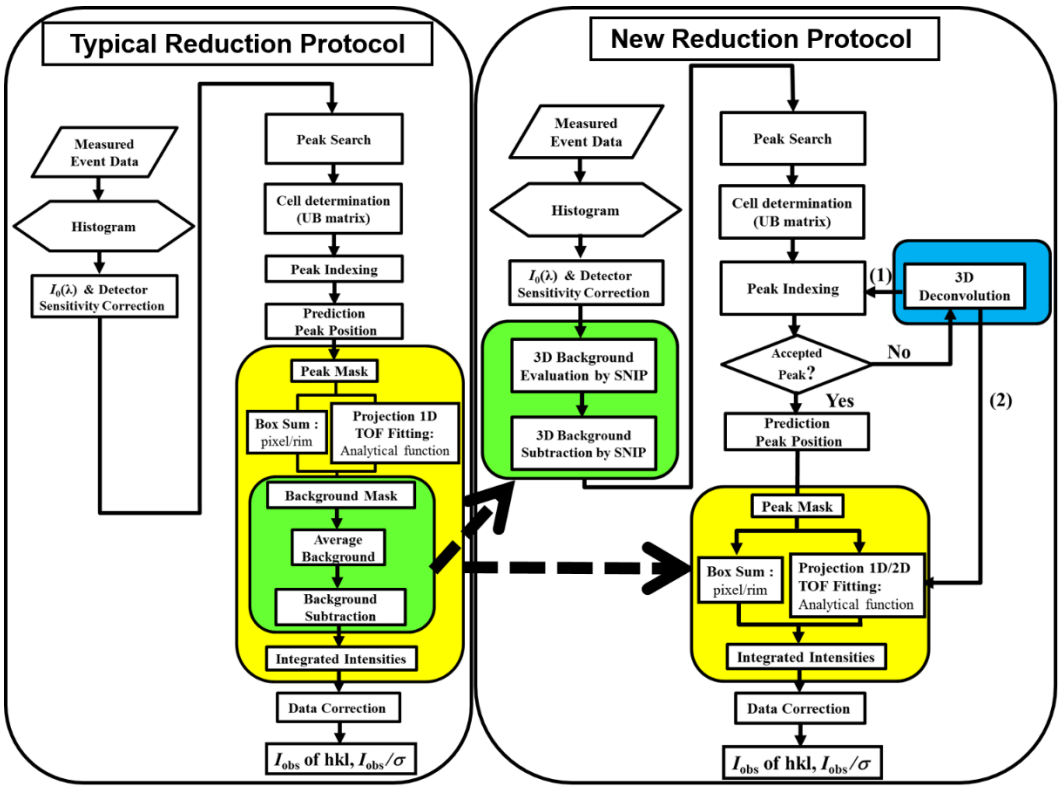

Fig 2. Data-flow of the data reduction protocol in TOF protein single-crystal neutron diffraction

The left panel shows a typical data reduction protocol for a TOF protein single-crystal neutron diffraction experiment. The right panel shows a proposed new reduction strategy for spectra obtained with a TOF protein single-crystal neutron diffractometer for crystals with large unit cells. The background evaluation routines in typical data reduction can be converted into a 3D background evaluation method by SNIP, as shown by the green colored components in the two protocols linked with a dashed arrow. The component for dealing with individual Bragg reflections is converted to the yellow component in the new reduction protocol. The blue colored component shows the routines for deconvoluting the overlapped Bragg reflections. The resolved peaks are re-indexed (1) and the fitting parameters including analytical functions thus obtained are also utilized in integration processes (2).

At the data correction stage in either a typical or the new reduction protocol, semi-empirical absorption corrections and multiple scattering must also be applied, using vanadium and sample scattering and inelasticity. A physics-related correction, such as the Lorentz factor, which is associated with the velocity to sweep the reciprocal space derived from the Laue neutron diffraction principle, must also be taken into account. These data are then further used in a least-squares refinement program to apply a variable wavelength extinction correction based on the Becker–Coppens formalism [18], normally using a Gaussian model with one variable parameter, the mosaic spread.

3. Summary

To integrate weak Bragg reflections from protein crystals with large unit cells using TOF protein single-crystal neutron diffractometry, we revisited the data reduction protocol, focusing on the background estimation to improve the signal-to-noise ratio. We introduced the high precise background estimation method at the initial stage of a new data reduction procedure to establish the precise evaluation of backgrounds under weak reflections from large unit-cell protein crystals (the green colored component in Fig. 2, right). In a typical data reduction protocol, the spot-based background estimation employed in the conventional data reduction protocol, as shown in Fig. 1, was applied for each observed Bragg reflection. This was locally averaged inside the peak domain (the green colored components Fig. 2, left). The procedure (the yellow colored component in Fig. 2, left) for each individual Bragg reflection involves spot-based background estimation. This procedure was divided into two independent components in the new data reduction protocol. With the background elimination method available for the background estimation of weak reflections and multidimensional space, authentic signals could be precisely extracted by analyzing the correlated backgrounds associated with X, Y, and TOF, indicating that this may prevent misclassification between these and the backgrounds. To overcome a serious peak-overlapping problem associated with large unit-cell protein crystals, we introduced a 3D deconvolution algorithm allowing precise integration of the peaks using a profile-fitting approach based on the known analytical shape of the reflections in the TOF directions, well understood from the characteristics of spallation neutron source (the blue colored component in Fig. 2, right).

A precise 4D summation integration or profile-fitting approach, as described in Fig. 1, remains to be realized to enable the data reduction of TOF protein single-crystal neutron diffraction data. The profile-fitting approach employed thus far in our studies is not intended for precise background estimation but for a precise profile shape in the TOF direction to improve the signal-to-noise ratio. Before realizing the profile-fitting approach for weak reflections, it will be necessary to establish precise background estimation and elimination methods. We will try to improve the signal-to-noise ratio more than 10% by employing software algorithm tailored to analyze the 4D spectra observed by Laue neutron diffraction experiment. Further modifications in the data reduction protocol for TOF protein single-crystal neutron diffractometry, which is designed for lattice constants greater than 250 Å, will be investigated.

4. Acknowledgement

This work was partly supported by JSPS KAKENHI Grant Number 15K04737.

References

[1] Tomoyori, K., Kurihara, K., Tamada, T. and Kuroki, R. (2015). *Journal of Physical society of Japan* DOI 036004 10.7566/JPSCP.8.036004

[2] Ohhara, T., Kusaka, K., Hosoya, T., Kurihara, K., Tomoyori, K., Niimura, N., Tanaka, I., Suzuki, J., Nakatani, T., Otomo, T., Matsuoka, S., Tomita, K., Nishimaki, Y., Ajima, T. and Ryufuku, S. (2009). *Nucl. Instr. Meth. Phys. Res., Sect.* A **600**, 195.

[3] Chatterjee, A., Mikkelson, R., Hammonds, J. and Worlton, T. (2002). *Appl. Phys*. A **S194**.
[4] Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr*. **26**.
[5] Otwinowski, Z. and Minor, W. (1997). *Methods Enzymol*. **276**, 307–326.
[6] Pflugrath, J. W. (1999). *Acta Cryst*. D **55**, 1718–1725.
[7] Kabsch, W. (1988). *J. Appl. Cryst*. **21**, 916–924.
[8] Busing, W. R. and Levy, H. A. (1967) *Acta Cryst* **22**, 457
[9] Jenkins, R. (1995). *Quantitative X-ray Spectrometry, CRC Press*, Boca Raton
[10] Van Grieken, R. and Markowicz, A. (2001). *Handbook of X-ray Spectrometry, CRC Press*, Boca Raton
[11] Clayton, E., Duerden, P. and Cohen, D. D. (1987). *Nucl Instr Meth Phys Res Sect* B **22** (1), 64-67
[12] Ryan, C. G., Clayton, E., Griffin, W. L., Sie, S. H. and Cousens, D. R. (1988). *Nucl Instr Meth Phys Res Sect* B **34**, 396-402
[13] Morhac, M. and Matousek, V. (2008). *Appl. Spectrosc* **62**, 91-106
[14] Tomoyori, K., Hirano, Y., Kurihara, K. and Tamada, T. (2015). *Journal of Physics* conference series **664** DOI:10.1088/1742-6596/664/7/072049
[15] Gutmann, M. J. (2005) SXD2001, ISIS Facility, Rutherford Appleton Laboratory, Oxford-shire, England
[16] Morhac, M., Matousek, V. and Kliman, J. (2003). *Digital Signal Processing*, **13**, No. 1, 144-171
[17] Tomoyori, K., Kusaka, K., Yamada, T., Hosoya, T., Ohhara, T., Kurihara, K., Tanaka, I., Katagiri, M. and Niimura, N. (2013). *Nucl Instr Meth Phys Res Sect* A **723**, 128-135
[18] Becker, P. and Coppens, P. (1974). *Acta Cryst* A **30**, 129 and 148.