Background elimination using the SNIP algorithm for Bragg reflections from a protein crystal measured by a TOF single-crystal neutron diffractometer

K Tomoyori, Y Hirano, K Kurihara and T Tamada
Quantum Beam Science Center, Japan Atomic Energy Agency,
2-4 Shirakata, Tokai, Ibaraki 319-1195, Japan

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

Abstract. In white neutron diffraction studies, the energy, intensity, and number of resolved Bragg peaks depend on detector resolution and background from physics processes. The statistics-sensitive non-linear iterative peak-clipping (SNIP) algorithm is widely used for subtracting the background under a peak. This paper reports the validation procedures and availability of the SNIP algorithm for time-of-flight (TOF) Bragg reflections observed with a TOF single-protein-crystal neutron diffractometer.

1. Introduction
In general, the characterization of Bragg peaks involves robust estimation of the underlying background. Several approaches have been proposed, ranging from simple estimation by analysis of the peaks’ side bands to spectrum fitting with analytical description of the background.

In spallation-neutron-source-based single-crystal diffractometers that use the time-of-flight (TOF) principle, convolution of “the time structure of the pulsed source” and “the neutron flight path” introduce significant asymmetries into the incident beam profile, which, when convolved with the diffraction profile of the sample, result in asymmetric diffraction peaks across the entire range of Bragg angles. We recently reported that the Landau and Vavilov distributions, which are used to describe the energy loss of charged particles traversing a thin absorber, were found to be in excellent agreement with the observed TOF profile [1, 2]. A Landau fit on the low-energy tail of the peak in addition to a Gaussian fit on the main peak region yields a closer approximation of irregular diffraction peak shapes. This closer approximation also allows for more accurate determination of the peak area.

In addition, the intensities of Bragg reflections from a protein crystal are comparatively weak, and the associated large unit cell parameters further aggravate this problem. Moreover, note that the weak Bragg reflections due to the large unit cells may be buried beneath the strong background resulting from incoherent scattering of hydrogen atoms, which approximately occupy one-half of all atoms in a protein molecule.

One important problem in the quantitative analysis of peak integration is the identification and quantification of the signal and the background. To that end, one needs to have reliable approximations of the background to estimate the signal above the background. When dealing with real spectra from experiments, non-iterative background approximation methods such as the three-window method [3] or polynomial fitting [4] are used. In addition, iterative methods such as simple...
multiple-point smoothing [5] or more sophisticated methods such as the statistics-sensitive non-linear iterative peak-clipping (SNIP) algorithm [6] are used.

The main advantage of the SNIP algorithm is its capability to cope with a large variety of background shapes. The SNIP algorithm is actually quite a popular algorithm for automated (or semi-automated) background subtraction. This algorithm, originally introduced for the treatment of proton induced x-ray emission [6], has been adapted for background elimination in coincidence \(\gamma\)-ray spectra [7]. This algorithm was also developed to account for spectra with poor and large statistics (extended dynamic range), search for a solution with the minimal number of parameters, and achieve full automation. Furthermore, Morhac [7-9] extended the SNIP algorithm to multidimensional spectra. In multidimensional spectra, the algorithm must recognize not only the continuous background but also include all combinations of coincidences of the background in some dimensions and peaks in the other ones. Using this algorithm for treating X–Y coincidence data collected by a wavelength shifting fiber (WLSF)-based neutron scintillation detector in TOF neutron protein crystallography may be advantageous.

We have already employed the SNIP algorithm for Bragg reflections from a protein crystal, which were observed with the IBARAKI biological crystal diffractometer (iBIX) installed in J-PARC/MLF [10], and confirmed its availability for background elimination [1, 2]. However, background estimation in the TOF direction remains inadequate to precisely determine the baseline because the width of the Bragg reflection profile varies with the TOF position, retaining its asymmetries associated with neutron production processes [11]. In addition, with background estimation in TOF neutron protein crystallography, accurately determining the baseline for the weak Bragg peaks in the wavelength-dependent, highly incoherent background can be difficult. In this study, we evaluated a suitable clipping window (CW) parameter required for accurately determining the baseline of Bragg reflections from a protein crystal in the TOF direction.

2. Principle of SNIP algorithm

The SNIP algorithm was introduced to separate useless information (i.e., background, noise, and detector artifacts) from useful information contained in the peak. The core procedure of the SNIP requires a pre-processing step called “the log square root (LLS) operator,” where the count \(y(i)\) in channel \(i\) is transformed into \(v(i)\) as follows:

\[
v(i) = \log\left(\log\left(\frac{y(i) + 1}{1} + 1\right) + 1\right).
\]

(1)

Usually the range of counts in a spectrum is very wide (e.g., \(1-10^6\) counts). Therefore, to compress the range of counts and enhance small peaks, the LLS operator is applied to the source spectrum.

In the algorithm, the background under the peak is evaluated in an iterative way (Fig. 1). The sampling interval is arbitrary and is called the clipping window (CW) \(M\). The CW \(M\) is the product of the bin width and number of iterations \(m\). For the \(m\)-th iteration, the content of the transformed bin \(v_m(i)\) is compared with the mean of the values at distances equal to \(\pm m\) and the updated spectrum is given as follows:

\[
v_{m+1}(i) = \min\left\{v_m(i), \frac{v_m(i-M) + v_m(i+M)}{2}\right\}.
\]

(2)
The process is repeated for every channel over \( m \) iterations. The estimated background \( B(i) \) is simply obtained from the final updated \( v_n(i) \) by applying the inverse LLS operator.

Morhac \textit{et al.} extended the SNIP algorithm to multidimensional data \cite{7-9} in such a way that the algorithm can recognize useless information (background and the combination of coincidences of the background with peak ridges) from useful information contained in \( n \)-fold coincidence peaks of an \( n \)-dimensional histogram.

3. Two-dimensional WLSF-type scintillation thermal neutron detector

The specifications required of neutron detectors used in a spallation neutron source facility are high-position resolution, high-detection efficiency, high counting rate, high \( n/\gamma \) ratio, etc. Especially, detectors characterized by high position and detection efficiencies are indispensable for crystallography diffractometers. WLSF neutron detectors have been employed in some diffractometers because they have the abovementioned specifications. The fibers in the WLSF neutron detectors were sandwiched between ZnS/\( ^{10}\text{B}_2\text{O}_3 \) scintillator screens. Neutrons absorbed by either the top or bottom scintillator screen cause scintillation light to be emitted owing to energy loss of the products of the nuclear reaction \( ^{10}\text{B} \ (n, \alpha) ^{7}\text{Li} \). These photons are reabsorbed by the WLSF. A fraction of the reemitted, wavelength-shifted photons in the WLSF reach photomultiplier tubes that convert the photons into an electric signal. Each fiber connects to an amplifier/discriminator, and the number of detected photoelectrons is counted by the signal-processing electronics hardware. The count is started based on a start time signal, which is a photonic digital pulse signal input first to a photon digital counter circuit, and conducted during the counting time, which corresponds to the coincidence time in the general coincidence circuit in field-programmable gate arrays (FPGAs). The photon integral value, which corresponds to two optical fibers, is obtained by using two X–Y coincidence circuits. When an integral value more than the photon discrimination value set beforehand is obtained from both circuits during the counting time, a neutron signal is output \cite{12, 13}.

4. Application of SNIP algorithm to TOF Bragg reflections from protein crystals

We performed diffraction experiments to evaluate the background using the SNIP algorithm in the TOF direction of NADH-cytochrome \( b_5 \) reductase (b5R) diffraction data collected using iBIX \cite{Y. Hirano \textit{et al.}, unpublished data). The b5R crystal belongs to the space group \( P2_1_2_1 \), with unit cell parameters \( a = 70.2 \), \( b = 83.1 \), and \( c = 47.9 \ \text{Å} \) \cite{14}. The diffraction data were collected using 30 detectors of the wavelength-shift-fiber type \cite{15}, each with a \( 133 \times 133 \ \text{mm}^2 \) sensitive area. To complete the data, datasets were collected using a wavelength range of 1.1–5.0 \( \text{Å} \) and a crystal-to-detector distance of 490 mm. The exposure times were 8.5 h for each set at 300 kW J-PARC accelerator power.

The raw event data acquired by iBIX DAQ \cite{16} were decoded into a 3D X–Y-TOF histogram domain specified by the TOF axis in addition to the \( x \) and \( y \) detector position coordinates using the ROOT/C++ analysis software \cite{17}. Fig. 2a shows 3D X–Y-TOF histogram, which is composed of 256 X, Y channel bins and 400 TOF bins between 11400 \( \mu \text{s} \) and 51400 \( \mu \text{s} \). The 3D X–Y-TOF histogram was projected onto the Y-TOF plane considering peak width in the X direction (Fig. 2b). The diffraction spots were lying on a straight line in the same direction as the Y-TOF plane. We introduced two graphical cut areas to avoid peak overlap on TOF projection, and eventually produced two types of 1D TOF histograms projected from the Y-TOF plane with the peak gate, as shown in Figs. 2c and 2d, respectively.

We have appropriately determined the peak CW widths in the TOF direction and estimated the 1D TOF background. It is clear that the lower the value of \( m \), the higher is the peak that will be subtracted, which implies that the underestimation of \( m \) leads to the degradation of peak efficiency. In contrast, overestimation of \( m \) leads to incomplete estimation of the background under peaks of interest. Therefore, a trade-off between peak efficiency and the amount of remaining background after subtraction determines the value of \( m \). In general, the value of \( m \) should depend on the peak width.
First of all, we applied the iteration parameter from $m = 1$ to 10. In Fig. 2c (gate I), the green baseline corresponding to $m = 6$ is suitable for peaks (4), (5), and (6) lying in the TOF region less than 27000 μs. In Fig. 2d (gate II), we employed $m = 6$ for peaks (8), (9), and (10). However, the value was not suitable for determining the baseline correctly for peak (7), and $m = 8$ was applied for estimating the background of peak (7).

On the other hands, the baseline is too small a CW to accurately estimate the baselines for peaks (1), (2), and (3) in Fig. 2c. Hence, we also tried the baseline estimation with increasing iteration parameter from $m = 6$ to 30 and determined appropriate iteration parameters for each existing peak, respectively. We determined optimum iteration number $m = 10$ for peaks (2) and (3), and $m = 11$ for the other broad peak (1).

After subtracting the background, we extracted Bragg peak signals and approximately applied the Gaussian for each peak and evaluated the pulse width, which depends on TOF position. Table 1 summarizes the fitting parameters of the peak position and its width, and the CW parameter of the SNIP algorithm thus obtained. Then, we investigated the relation between CW width and peak width.
5. Discussion

We have applied a clipping operation-based algorithm with an automatically adjustable window to the widths of identified peak regions corresponding to experimental data [1]. However, the width of Bragg reflections in the TOF direction becomes considerably broader as TOF position increases; this is attributed to the neutron moderation process associated with decreasing neutron energy [11]. The estimation of one correct CW width value valid for all the spectra over the entire TOF axis range is impossible. Therefore, in the present work, we studied the optimum CW in accordance with the TOF position. According to Table 1, a peak CW equivalent to around 1.5 times the FWHM is an accurate localized approximation. Regarding the much higher TOF region, the applicable peak CW width might be deliberately chosen because of the much broader peak width with its asymmetric tails.

In addition, the LLS operator could be effective for improving the accuracy of background evaluation on small Bragg reflections from a protein crystal. In fact, some small peaks in Fig. 2c and 2d are clearly identified and the baseline is correctly estimated, demonstrating the feasibility of precise integration of small peaks from a protein crystal with large unit cells.

Further progress depends on the ability to correctly analyze the coincidence data in three-dimensional space, including TOF axis in addition to x, y detector coordinates. Needless degradation of dimension before background elimination could lead to loss of meaningful information about the background. As usual, the boundary between the background and the signal region is determined in terms of x, y, TOF; then, the original dimension is sometimes degraded into one-dimensional TOF direction to integrate peaks. In contrast, the three-dimensional spectra of X–Y-TOF include the effects of all combinations of coincidences of the background in the X direction and the peaks in the Y direction and vice versa in addition to a continuous background. The SNIP algorithm can realize background evaluation without decreasing the original dimension, thus implying the feasibility of

### Table 1. Estimated fitting parameters and CW width on some peaks

| Peak Index | Center<sup>a,b</sup> (µs) | Width<sup>a,b</sup> (µs) | Estimated Iteration Number | CW<sup>c</sup> (µs) | CW / Width<sub>FWHM</sub><sup>d</sup> |
|------------|------------------------|----------------------|-------------------------|----------------|------------------------|
| 1          | 4.31 × 10<sup>4</sup> ± 4.72 | 3.48 × 10<sup>2</sup> ± 2.58 | 11                      | 1100           | 1.4                    |
| 2          | 3.54 × 10<sup>4</sup> ± 6.84 | 2.93 × 10<sup>2</sup> ± 3.65 | 10                      | 1000           | 1.5                    |
| 3          | 2.99 × 10<sup>4</sup> ± 3.96 | 2.85 × 10<sup>2</sup> ± 2.64 | 10                      | 1000           | 1.5                    |
| 4          | 2.60 × 10<sup>4</sup> ± 6.29 | 1.83 × 10<sup>2</sup> ± 3.27 | 6                       | 600            | 1.4                    |
| 5          | 2.29 × 10<sup>4</sup> ± 6.44 | 1.57 × 10<sup>2</sup> ± 3.77 | 6                       | 600            | 1.6                    |
| 6          | 2.05 × 10<sup>4</sup> ± 19.3 | 2.12 × 10<sup>2</sup> ± 17.1 | 6                       | 600            | 1.2                    |
| 7          | 2.75 × 10<sup>4</sup> ± 3.85 | 2.48 × 10<sup>2</sup> ± 2.64 | 8                       | 800            | 1.4                    |
| 8          | 2.42 × 10<sup>4</sup> ± 9.76 | 2.01 × 10<sup>2</sup> ± 5.53 | 6                       | 600            | 1.3                    |
| 9          | 2.16 × 10<sup>4</sup> ± 7.79 | 1.90 × 10<sup>2</sup> ± 5.19 | 6                       | 600            | 1.4                    |
| 10         | 1.94 × 10<sup>4</sup> ± 8.37 | 1.72 × 10<sup>2</sup> ± 5.94 | 6                       | 600            | 1.5                    |

---

<sup>a</sup> Peak center and width (standard deviation) were evaluated applying Gaussian to each peak after subtracting background.

<sup>b</sup> Error values shown are fitting error estimates in least square analysis.

<sup>c</sup> CW<sub>M</sub> is the product of bin width (100 µs) and number of iterations <i>m</i>.

<sup>d</sup> Width<sub>FWHM</sub> is derived from 2.354 x Width, given that FWHM = 2.354 σ.
estimating cross ridges associated with X–Y coincidence neutron events detected by a WLSF neutron detector.

In FPGA-based online analysis, we cannot discriminate whether each scattered neutron event is derived from coherent or incoherent properties because each event depends on the scattering probability of atoms inside a protein crystal. However, by employing this algorithm, one can include the linear background as well as both ridges near a peak into a created base-volume background spectrum. After subtracting the background spectrum from original spectrum, we can get pure X–Y-TOF 3D peaks. If we want to create one- or two-dimensional slices from a three-dimensional spectrum, we first have to estimate and subtract the background in three-dimensional space using the three-dimensional SNIP algorithm [7-9]. Then, we can generate slices, fit, and integrate peaks.

6. Conclusion

We evaluated the peak CW required for TOF background estimation of Bragg reflections observed in a spallation-based single-crystal neutron diffraction experiment. The SNIP algorithm identified equally well the weak TOF Bragg reflections as well as strong reflections with variable peak width and high-level background. The CW parameter should be variable with the peak width in the TOF spectrum and a peak CW equivalent to around 1.5 times the FWHM may be required to correctly estimate the baseline. The CW parameter should be optimized deliberately for the broader peaks with asymmetries in the higher end of the TOF spectrum. We confirmed that the SNIP algorithm is suitable for determining the baseline of Bragg reflections in the TOF direction.

7. Acknowledgement

We sincerely appreciate all the colleagues of iBIX core member who worked with us for this experiment. We are also deeply indebted to Prof. Kunio Miki who provided neutron diffraction data of NADH-cytochrome b₅ reductase.

References

[1] 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 vol 723 pp 128-135
[2] Tomoyori K, Kusaka K, Yamada T and Tamada T 2014 J Struct Funct Genomics vol 15 pp 131-135
[3] R. Jenkins, Quantitative X-ray Spectrometry, CRC Press, Boca Raton, 1995
[4] R. van Grieken and A. Markowicz, Handbook of X-ray Spectrometry, CRC Press, Boca Raton, 2001
[5] Clayton E, Duerden P and Cohen DD, A discussion of PIXAN and PIXANPC: The Aaec PIXE analysis computer packages, Nucl Instr Meth Phys Res Sect B: Beam Interact. Matter. Atoms 1987 vol 22 (1) pp 64-67
[6] Ryan CG, Clayton E, Griffin WL, Sie SH and Cousens DR 1988 Nucl Instr Meth Phys Res Sect B vol 34 pp 396-402
[7] Morhac M, Klimma J, Matousek V, Veselsky M and Turzo I 1997 Nucl Instr Meth Phys Res Sect A, 401 pp113-132
[8] Morhac M and Matousek V 2008, Appl. Spectrosc vol 62, pp 91-106
[9] Morhac M 2009, Nucl Instr Meth Phys Res Sect A vol 600 pp 478-487
[10] Tanaka I, Kusaka K, Hosoya T, Niimura N, Ohhara T, Kurihara K, Yamada T, Ohnishi Y, Tomoyori K and Yokoyama T 2010 Acta Cryst. D vol 66 pp1194-1197
[11] Ikeda S and J. M. Carpenter 1985 Nucl Instr Meth Phys Res Sect A vol 239 pp536-544
[12] Katagiri M, Nakamura T, Ebine M, Birumachi A, Sato S and E.M. Shoondeveld 2007 Nucl Instr Meth Phys Res Sect A vol 573 pp 149-152
[13] Nakamura T, Kawasaki T, Hosoya T, Toh K, Oikawa K, Sakasai K, Ebine M, Birumachi A, Soyama K and Katagiri M 2012 Nucl Instr Meth Phys Res Sect A 686 pp 64-70
[14] Yamada M, Tamada T, Takeda K, Matsumoto F, Ohno H, Kosugi M, Takaba K, Shoyama Y, Kimura S, Kuroki R and Miki K 2013 *Journal of Molecular Biology* vol 425 pp 4295-4306

[15] Kusaka K, Hosoya T, Yamada T, Tomoyori K, Ohhara T, Katagiri M, Kurihara K, Tanaka I and Niimura N 2013 *J Synchrotron Radiat* 20 pp 994-998

[16] Hosoya T, Nakamura T, Katagiri M, Birumachi A, Ebine M and Soyama K 2009 *Nucl Instr Methods Phys Res Sect A* vol 600 pp 217-219

[17] R. Brun and F. Rademakers 1997 *Nucl Instr Meth Phys Res Sect A* vol 389 pp 81-86