https://doi.org/10.2352/ISSN.2470-1173.2017.17.COIMG-415

Synchrotron X-Ray Diffraction Dynamic Sampling for Protein Crystal Centering

Nicole M. Scarborough[a], G. M. Dilshan P. Godaliyadda[b], Dong Hye Ye[b], David J. Kissick[c], Shijie Zhang[a], Justin A. Newman[a], Michael J. Sheedlo[a], Azhad Chowdhury[a], Robert F. Fischetti[c], Chittaranjan Das[a], Gregery T. Buzzard[d], Charles A. Bouman[b] and Garth J. Simpson[a]*

[a]Department of Chemistry, Purdue University, West Lafayette, IN, 47907
[b]Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 47907
[c]GM/CA@APS, X-ray Science Division, Argonne National Laboratory, Lemont, IL, 60439
[d]Department of Mathematics, Purdue University, West Lafayette, IN, 47907

Correspondence email: gsimpson@purdue.edu

Abstract

A supervised learning approach for dynamic sampling (SLADS) was developed to reduce X-ray exposure prior to data collection in protein structure determination. Implementation of this algorithm allowed reduction of the X-ray dose to the central core of the crystal by up to 20-fold compared to current raster scanning approaches. This dose reduction corresponds directly to a reduction on X-ray damage to the protein crystals prior to data collection for structure determination. Implementation at a beamline at Argonne National Laboratory suggests promise for the use of the SLADS approach to aid in the analysis of X-ray labile crystals. The potential benefits match a growing need for improvements in automated approaches for microcrystal positioning.

Background and Motivation

Beam-scanning or sample-scanning measurements are commonplace in many imaging application, spanning wavelengths from the visible in confocal microscopy to X-ray in scanning electronic microscopy. In many instances, the time required to produce a signal at a particular location represents the rate-limiting step for image generation. When the single-pixel measurement time is much less than the random-access time, opportunities for dynamic sampling emerge, in which the location of the next measurement is informed by the preceding set of measurements rather than chosen blindly.

X-ray diffraction imaging (XRDI) represents precisely such a scenario. XRDI is routinely used in structural biology as a means of automatically positioning protein crystals prior to diffraction data collection at synchrotron sources.[1-5] In XRDI, a full X-ray scattering pattern is collected at each location in a sample using a tightly collimated X-ray source. In the presence of a protein crystal, a series of sharp diffraction features appear in the scattering image. Peak identification is used to indicate locations corresponding to crystal locations. In this approach, a raw high resolution image of X-ray scattering is converted into a single scalar value in an image of crystal position.

While XRDI has the distinct advantage of generating image contrast based on the measurement of highest priority (i.e., diffraction efficiency), it suffers from two key limitations. First, XRDI is relatively slow, since a high resolution (e.g., 4 or 16 MPixel) image is acquired for each point in the real-space image. Integration of 0.5-2 seconds per pixel is often used to generate sufficiently high signals for reliable crystal positioning given the relatively low scattering cross-sections of the atoms within organic molecules. Second, XRDI can result in significant sample damage from X-ray exposure prior to data collection.

Damage from X-ray exposure represents one of the most challenging issues to address in protein structure determination by X-ray diffraction. For every 1 diffracted X-ray photon contributing to structure determination, roughly 10 are absorbed to produce local disorder and loss of diffracted power.[6,7] The cumulative damaging effects of X-ray absorption ultimately dictate the signal to noise ratio (SNR) achievable in a given diffraction experiment, and the corresponding confidence in the resulting protein structures produced by the analysis. As such, major efforts have evolved to reduce SNR losses associated with damage, such as using high flux synchrotron X-ray sources and performing diffraction measurements under cryogenic conditions. At cryogenic temperatures, diffusion of damaging radicals produced upon X-ray exposure is greatly suppressed.[6,8,9] However, it is not always clear whether the diffraction results obtained at cryogenic temperatures are representative of the structures present under ambient conditions. These advances have significantly pushed the boundaries of the measurements to analysis of ever smaller crystals using ever smaller X-ray beams, such that diffraction experiments can now be performed on crystals smaller than 1 μm in dimension.

This size reduction has in turn exacerbated practical challenges in positioning protein crystals within the X-ray beam prior to diffraction data collection. For small crystals requiring high flux for detection and for X-ray labile crystals, this additional exposure can reduce the remaining dose available for diffraction data collection. As the sizes of both the sources and the crystals continue to be reduced, the challenges of centering on the basis of diffraction will continue to grow in significance.

In this work, exposures to the central core of protein crystals are greatly reduced using dynamic scan patterns, in which the location for the next diffraction measurement is selected on the basis of the preceding set of measurements. In this manner, the total number of locations within the field of view exposed to X-rays is reduced by up to 20-fold compared to conventional raster scanning. Implementation of dynamic sampling on a synchrotron beamline at Argonne National Laboratory suggests direct compatibility of this approach with hardware currently in place at many macromolecular synchrotron diffraction facilities.

Theoretical Foundation

The theoretical background underpinning SLADS illustrated in Figure 1 is detailed elsewhere[10] and briefly summarized below. For a ground truth underlying object X consisting of N pixels, the set of k measurements at locations S combine to generate the set of known information Y.

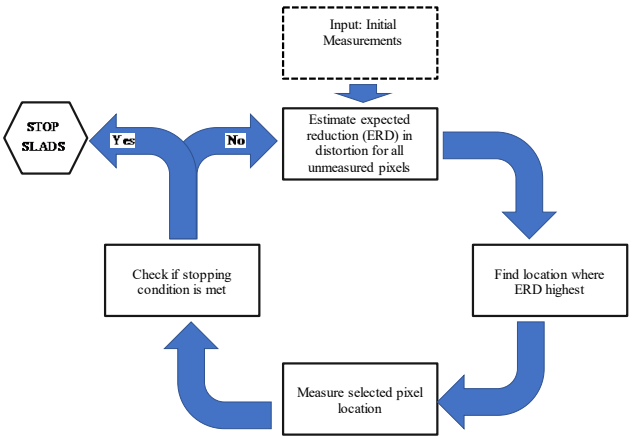

Figure 1. Schematic overview of the SLADS reconstruction algorithm.

$$Y^{(k)} = \begin{bmatrix} s^{(1)}, X_{s^{(1)}} \\ \vdots \\ s^{(k)}, X_{s^{(k)}} \end{bmatrix} \quad (1)$$

The primary goal of the SLADS algorithm is to identify the location $s^{(k+1)}$ that reduces the subsequent reconstruction distortion between the ground truth and recovered images X and $\hat{X}^{(k)}$, respectively. For a binary image, the distortion D for pixel r is defined by the following.

$$D\left(X_r, \hat{X}_r\right) = \begin{cases} 0 \text{ if } X_r = \hat{X}_r \\ 1 \text{ if } X_r \neq \hat{X}_r \end{cases} \quad (2)$$

Since increasing the number of measurements will generally improve the accuracy of the reconstruction and reduce the distortion, the reduction in distortion R from measurement of the s pixel after k preceding measurements is given by the following definition.

$$R^{(k;s)} = \sum_r D\left(X_r, \hat{X}_r^{(k)}\right) - D\left(X_r, \hat{X}_r^{(k;s)}\right) \quad (3)$$

In practice, X is not known in advance. However, the expected reduction in distortion (ERD = $\bar{R}$) can be estimated from the expectation value of R.

$$\bar{R}^{(k;s)} = E\left[\sum_r D\left(X_r, \hat{X}_r^{(k)}\right) - D\left(X_r, \hat{X}_r^{(k;s)}\right) \middle| Y^{(k)}\right] \quad (4)$$

The sequentially optimized sampling location for the k+1 measurement is the position that maximizes the expected reduction in distortion from Eq. 1.4. In SLADS, the relationship between the measurements Y and the ERD is a regression function informed by an offline training process.

$$R^{(s)} \approx \sum_{r \in \Omega} h_r^{(s)} D\left(X_r, \hat{X}_r\right) \quad (5)$$

$$h_r^{(s)} = \exp\left\{-\frac{c}{2\left(\sigma^{(s)}\right)^2}\|r-s\|^2\right\} \quad (6)$$

In Eqs. 5 and 6, $\sigma^{(s)} = \min_{t \in S}\left\{\|s-t\|^2\right\}$, where, Ω is the set of all locations in the image, and S is the set of sampled locations.

Experimental Methods

Full length mCherry was cloned into pGEX6P1 and transformed into Rosetta cells using standard cloning protocols detailed previously. The crystals used in these studies were grown using both sitting drop and hanging drop vapor diffusion methods in mother liquor containing 100 mM Tris pH 8.0, 100 mM sodium acetate and 30% PEG 4000 at room temperature, as has been previously reported [11]. The crystals grew to large clusters of rod-shaped crystals over the course of 1-4 days.

Contrast in the XRDI images corresponds to locations of protein-like diffraction, which was assessed using the program DISTL [12]. DISTL is a part of the package, LABELIT [13], which estimates potential Bragg candidates. In three steps: i) isolating diffraction like peaks from the background in a diffraction image considering the noise variability in local environment, ii) validating the isolated peaks from the rejection of possible sources of ice-rings, salt particles or crystal disorder, and iii) gauging size and shapes of each peak. DISTL estimates diffraction peaks more quickly than full-blown indexing and processing of diffraction data (normally performed with programs like XDS [14,15], MOSFLM [16], HKL2000 [17]., in which patterns containing sufficient numbers of identified peaks were classified as corresponding to protein crystals, while those below the threshold were classified as blank (solution or non-crystalline protein).

Results and Discussion

Prior to implementation of SLADS, a significant effort was devoted to assessing the advantages and limitations in reconstructions in which ground truths were known in advance. Ground-truth data were obtained by performing a full XRDI acquisition with diffraction performed at each location. The ground truth data could then be compared directly with the results of the reconstructions for statistical evaluation of performance. A representative X-ray scattering pattern recorded from a single confined 5μm diameter region is shown in Figure 2. The bright puncta in the scattering pattern correspond to diffraction-like peaks. Using peak-identification algorithms detailed in the Experimental Methods section, the scattering pattern shown in Figure 2 was reduced to a single binary classification: 1 = protein crystal, 0 = background.

Prior to implementation of SLADS for protein crystal positioning at a synchrotron facility, simulations were performed using known ground-truth diffraction images and surrogates.

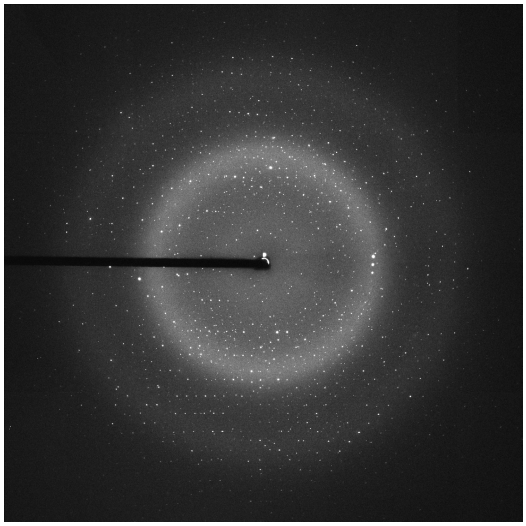

Figure 2. Representative X-ray scattering pattern containing protein-like diffraction features (bright puncta) recorded from a single pixel in the XRDI. The black shadow to the left and center is cast from beam-stop to remove the primary specular X-ray beam. The spot count serves as a classifier for assigning locations as either protein crystals or background.

A comparison of the ground truth diffraction-based mapping results and the recovered binary map is shown in Figure 3. In the diffraction mapping, a 4Mpixel X-ray scattering pattern such as shown in Figure 2 was acquired at each spatial position in the image and the number of diffraction-like peaks used as the greyscale value in the corresponding diffraction images. A binary map was generated upon thresholding the diffraction image. Dynamic sampling of the ground truth results allowed image reconstruction and assessment of reconstruction error.

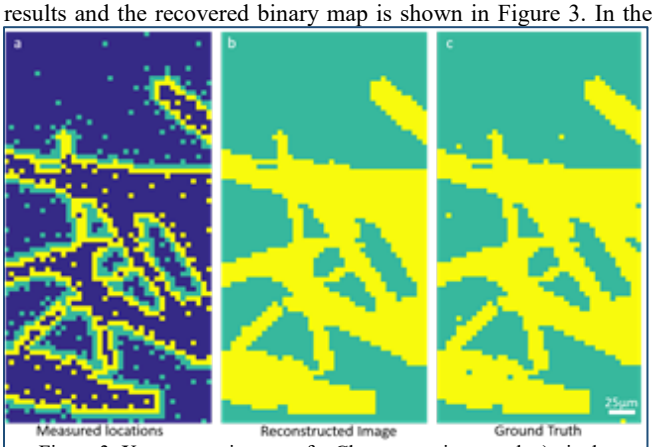

Figure 3. X-ray raster images of mCherry protein crystal: a) pixels accessed by dynamic sampling, with yellow indicating crystalline regions, teal noncrystalline locations, and dark blue unsampled pixels, b) the corresponding reconstructed image of crystal locations, and c) the ground truth diffraction image expressed as a binary map.

Plots of the reconstruction error arising from the simulation results shown in Figure 3 and analogous simulations performed with a 1 μm beam allowed assessment of the capabilities of the SLADS approach as the number of pixels is increased. The total distortion given by Equation 2 was normalized and plotted for both SLADS and a more conventional low-discrepency sampling approach. In both instances, SLADS offered marked improvements in performance. Notably, the differential was significantly greater for higher resolution measurements, in which a greater total fraction of pixels were in-painted in the reconstructions. These results suggest that reasonably accureate reconstructions can be obtained with only a few % of the pixels sampled for the crystals considered in the present work.

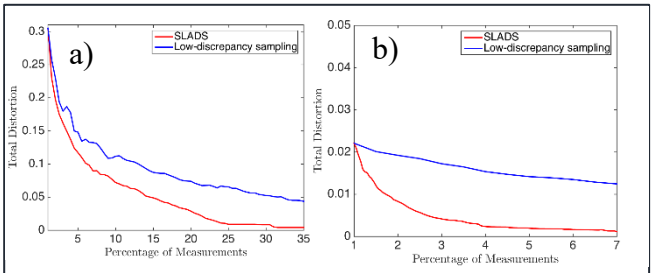

Figure 4. Comparison of the reconstruction distortion produced by SLADS and conventional low discrepancy sampling for the measurements obtained with a 5 μm diameter X-ray source (a) and the surrogate TPEF measurements for a 1 μm diameter beam (b).

From the reconstructions in Figure 3, the SLADS algorithm preferentially samples pixels at the edges between regions of different classification. The origin of this bias is straightforward; the ambiguity for classification is greatest for the pixels on the borders, such that the overall reduction in normalized distortion (ND) is optimized by preferentially sampling those pixels most likely to be potentially mis-classified.

An analogous set of operations was also performed for two-photon excited ultraviolet fluorescence (TPE-UVF) images, which served as surrogates for the next generation of X-ray "minibeam" sources targeting ~1 μm diameter beams.

Following the proof-of-concept studies, SLADS was implemented on a beamline with the GM/CA@APS group at Argonne National Laboratory, the results of which are summarized

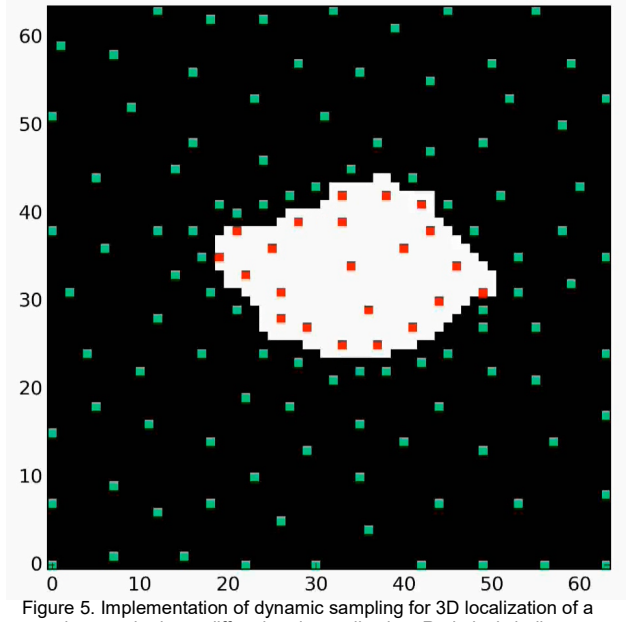

Figure 5. Implementation of dynamic sampling for 3D localization of a protein crystal prior to diffraction data collection. Red pixels indicate locations of protein-like diffraction and teal the absence. The reconstructed crystal is indicated in white.

in Figure 5. Prior to data collection for structure determination, protein crystals are typically centered in X and Y, then rotated 90º and centered in X and Z, consistent with the combined data set in Figure 4. Once centered, diffraction measurements are typically recorded every few degrees as the crystal is rotated through a pre-set range of angles. Centering prior to data collection ensures that the protein remains within the beam during this rotation operation.

Consistent with the predicted trends based on the simulations, SLADS reliably identified locations of the protein crystals with significant reductions in exposure to damaging X-rays prior to data collection. Specifically, the crystals were positioned with just 3% of the locations and just 5% of the protein exposed to X-rays. Furthermore, the SLADS algorithm preferentially samples the crystal edges, leaving unexposed the central regions of the protein crystal that produce the brightest diffraction.

Acknowledgements

NMS, SZ, JAN, MJS, AC, CD, and GJS gratefully acknowledge support from the NIH Grant Numbers R01GM-103401 and R01GM-103910 from the NIGMS.

References

1 Cherezov, V. *et al.* Rastering strategy for screening and centring of microcrystal samples of human membrane proteins with a sub-10 µm size X-ray synchrotron beam. *J. R. Soc., Interface* **6**, S587-S597, doi:10.1098/rsif.2009.0142.focus (2009).

2 Hilgart, M. C. *et al.* Automated sample-scanning methods for radiation damage mitigation and diffraction-based centering of macromolecular crystals. *J. Synchrotron Radiat.* **18**, 717-722, doi:10.1107/s0909049511029918 (2011).

3 Song, J. *et al.* Diffraction-based automated crystal centering. *J. Synchrotron Radiat.* **14**, 191-195, doi:doi:10.1107/S0909049507004803 (2007).

4 Stepanov, S. *et al.* Fast fluorescence techniques for crystallography beamlines. *J. Appl. Cryst.* **44**, 772-778, doi:10.1107/s0021889811016748 (2011).

5 Aishima, J. *et al.* High-speed crystal detection and characterization using a fast-readout detector. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **66**, 1032-1035, doi:doi:10.1107/S0907444910028192 (2010).

6 Holton, J. M. A beginner's guide to radiation damage. *J Synchrotron Radiat* **16**, 133-142, doi:10.1107/s0909049509004361 (2009).

7 Garman, E. Radiation damage in macromolecular crystallography: what is it and why should we care? *Acta Crystallogr. Sect. D* **66**, 339-351, doi:doi:10.1107/S0907444910008656 (2010).

8 Burmeister, W. Structural changes in a cryo-cooled protein crystal owing to radiation damage. *Acta Crystallogr. Sect. D* **56**, 328-341, doi:doi:10.1107/S0907444999016261 (2000).

9 Nave, C. & Garman, E. F. Towards an understanding of radiation damage in cryocooled macromolecular crystals. *J. Synchrotron Radiat.* **12**, 257-260, doi:10.1107/s0909049505007132 (2005).

10 Godaliyadda, G. M. D. *et al.* A Supervised Learning Approach for Dynamic Sampling. *Electronic Imaging* **2016**, 1-8 (2016).

11 Shu, X., Shaner, N. C., Yarbrough, C. A., Tsien, R. Y. & Remington, S. J. Novel chromophores and buried charges control color in mFruits. *Biochemistry* **45**, 9639-9647, doi:10.1021/bi0607731 (2006).

12 Zhang, Z., Sauter, N. K., van den Bedem, H., Snell, G. & Deacon, A. M. Automated diffraction image analysis and spot searching for high-throughput crystal screening. *Journal of applied crystallography* **39**, 112-119 (2006).

13 Sauter, N. K., Grosse-Kunstleve, R. W. & Adams, P. D. Robust indexing for automatic data collection. *Journal of applied crystallography* **37**, 399-409 (2004).

14 Diederichs, K. Some aspects of quantitative analysis and correction of radiation damage. *Acta Crystallographica Section D: Biological Crystallography* **62**, 96-101 (2006).

15 Kabsch, W. Xds. *Acta Crystallographica Section D: Biological Crystallography* **66**, 125-132 (2010).

16 Leslie, A. G. & Powell, H. R. in *Evolving methods for macromolecular crystallography* 41-51 (Springer, 2007).

17 Minor, W., Tomchick, D. & Otwinowski, Z. Strategies for macromolecular synchrotron crystallography. *Structure* **8**, R105-R110 (2000).

Author Biography

The authors represent a multi-disciplinary team with expertise spanning ultrafast optics, image reconstruction algorithm development, and protein structure determination. Many advances routinely adopted worldwide for protein structure determination were initially demonstrated at GM/CA@APS, including the development of X-ray raster scanning and mini-beam diffraction. The Purdue team spans expertise in Engineering, Mathematics, and Chemistry.

The presenting author Garth J. Simpson received his Ph.D. in Chemistry from the University of Colorado at Boulder in 2000, and started as a faculty member at Purdue University in 2001.He is a co-inventor on the use of nonlinear optical imaging methods for protein crystal detection and has long-standing collaborations with GM/CA@APS to implement these capabilities at synchrotron facilities.