Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 25 August 2010
Accepted 5 December 2010

Structure-factor analysis of femtosecond microdiffraction patterns from protein nanocrystals

Richard A. Kirian,[a] Thomas A. White,[b] James M. Holton,[c,d] Henry N. Chapman,[b,e] Petra Fromme,[f] Anton Barty,[b] Lukas Lomb,[g] Andrew Aquila,[b] Filipe R. N. C. Maia,[h] Andrew V. Martin,[b] Raimund Fromme,[f] Xiaoyu Wang,[a] Mark S. Hunter,[f] Kevin E. Schmidt[a] and John C. H. Spence[a]*

[a]Department of Physics, Arizona State University, Tempe, Arizona 85287, USA, [b]Center for Free-Electron Laser Science, DESY, Notkestrasse 85, 22607 Hamburg, Germany, [c]Advanced Light Source, Lawrence Berkeley Laboratory, Berkeley, CA 94720, USA, [d]Department of Biochemistry, University of California, San Francisco, CA 945158-2330, USA, [e]University of Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany, [f]Department of Biochemistry, Arizona State University, Tempe, Arizona 85287, USA, [g]Max Planck Institute for Medical Research, Jahnstrasse 29, 69120 Heidelberg, Germany, and [h]Department of Cell and Molecular Biology, Laboratory of Molecular Biophysics, Uppsala University, Husargatan 3 (Box 596), SE-751 24 Uppsala, Sweden. Correspondence e-mail: spence@asu.edu

A complete set of structure factors has been extracted from hundreds of thousands of femtosecond single-shot X-ray microdiffraction patterns taken from randomly oriented nanocrystals. The method of Monte Carlo integration over crystallite size and orientation was applied to experimental data from Photosystem I nanocrystals. This arrives at structure factors from many partial reflections without prior knowledge of the particle-size distribution. The data were collected at the Linac Coherent Light Source (the first hard-X-ray laser user facility), to which was fitted a hydrated protein nanocrystal injector jet, according to the method of serial crystallography. The data are single 'still' diffraction snapshots, each from a different nanocrystal with sizes ranging between 100 nm and 2 μm, so the angular width of Bragg peaks was dominated by crystal-size effects. These results were compared with single-crystal data recorded from large crystals of Photosystem I at the Advanced Light Source and the quality of the data was found to be similar. The implications for improving the efficiency of data collection by allowing the use of very small crystals, for radiation-damage reduction and for time-resolved diffraction studies at room temperature are discussed.

1. Introduction

With the recent invention of the X-ray laser, an opportunity has arisen to break the nexus between radiation dose and spatial resolution (described by Howells *et al.*, 2009). It has been proposed that femtosecond X-ray pulses can be used to out-run damage processes, when using single pulses so brief that they terminate before the onset of significant radiation damage (Neutze *et al.*, 2000; Solem, 1986). A recent review of this approach for diffractive imaging, holography and crystallography is given elsewhere (Chapman, 2009). Experiments at the FLASH vacuum ultraviolet free-electron laser (FEL) confirmed this idea at resolution lengths greater than 6 nm (Chapman *et al.*, 2006). However, until now, it had not been demonstrated that quantitative high-quality diffraction data can be extracted from the scattering of intense femtosecond X-ray pulses focused onto a protein nanocrystal or single particle, which would greatly enhance the prospects for structure determination using submicron crystals and for time-resolved crystallography.

Recently, the first femtosecond protein nanocrystallography diffraction experiments were carried out at the Linac Coherent Light Source (LCLS) X-ray free-electron laser (XFEL) (Chapman *et al.*, 2011). Data for this analysis were provided by the CFEL/ASG/ASU collaboration described elsewhere (Chapman *et al.*, 2011). In the experimental setup at the 'CFEL–ASG multi-purpose' (CAMP) end-station (Strüder *et al.*, 2010) on the Atomic, Molecular and Optical Science (AMO) beamline (Bozek, 2009), single-crystal diffraction data were collected from a stream of nanocrystals carried in a continuous liquid water jet that flows, in vacuum, across the focused LCLS X-ray beam, according to the method of serial crystallography (Spence & Doak, 2004). Experimental details are given in §4. Since the 'exposure time' for these patterns is typically a few tens of femtoseconds, these patterns are randomly oriented 'stills' (Rossmann & Erickson, 1983), which provide only a partial integral over the reciprocal-space volume of each Bragg spot in contrast to conventional protein crystallography. As shown in Fig. 1 (see also the experimental patterns in Chapman *et al.*, 2011), for the smallest crystals consisting of just a few unit cells (Hunter *et al.*, 2011), there are diffraction maxima, in the form of interference fringes, extending entirely from one Bragg peak to the next [the basic theory is described by Woolfson (1997), ch. 2]. Crystal-size effects (Scherrer broadening) thus dominate the size of the Bragg spots, rather than beam divergence, energy spread or mosaicity, as in conventional protein crystallography. For this type of data conventional oscillation data-processing methods (Arndt & Wonacott, 1977) are not applicable because there is no *a priori* way to tell if a spot is a 'major' Bragg peak or merely a subsidiary maximum, let alone estimate its partiality. However, once reasonably accurate structure factors and crystal orientations are obtained, it may be possible to extract effective partialities using a post-refinement procedure similar to that described by Rossmann *et al.* (1979) for oscillation cameras, but this was not done in this work. Calculating partialities for spots from laser-illuminated nanocrystals will be much more challenging because a small sphere in reciprocal space is a poor model of the 'rocking curve' in this case. Specifically, the Fourier transform of the crystal shape ('shape transform') is convoluted with each reciprocal-lattice point, spreading the total intensity into subsidiary maxima that may range all over reciprocal space, as first shown by von Laue (1936). In fact, since the observation of a 'spot' does not necessarily imply that the main Bragg peak intersected the Ewald sphere, these subsidiary maxima create a number of challenges for auto-indexing. In addition, all of the data analysis must be highly automated, since human evaluation of this large number of patterns is not practical.

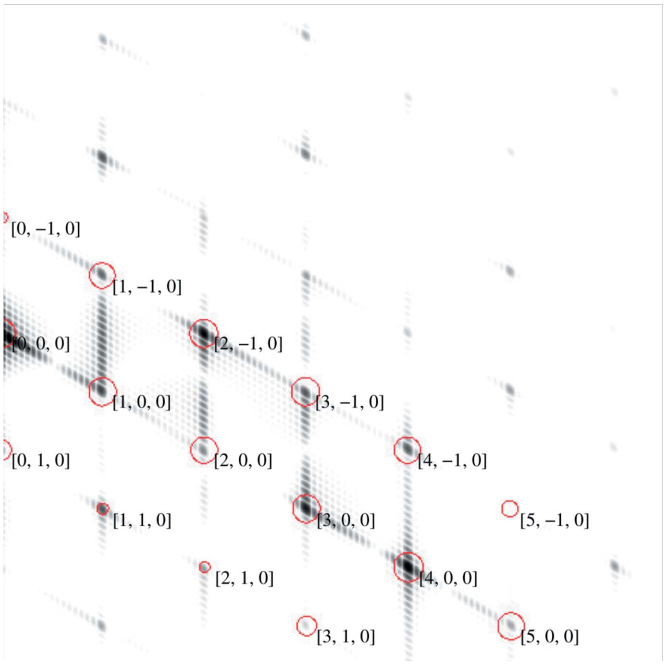


Figure 1
Simulation of an ideal 1.8 keV diffraction pattern for a perfect (no mosaicity) Photosystem I nanocrystal with 17 × 17 × 30 unit cells along the *a*, *b* and *c* directions (~0.5 μm in size), with a spectral width of 0.1% and beam divergence of 1.5 mrad. Red rings indicate the intersection of the Ewald sphere with the boundary of the integration domains described in §2, for a value of $\delta = 5\ \mu\text{m}^{-1}$. Fringes result from interference between shape transforms, and this effect dominates beam divergence and energy spread in the beam. The crystal is in a random orientation, near the *hk*0 projection.

In this paper we are concerned with the extraction of structure factors from snapshot diffraction patterns, each from a different nanocrystal. An approach to this problem was published recently based on simulations (Kirian *et al.*, 2010), which we use as a guide here for the processing of experimental data from the LCLS. A suite of crystallography software was developed, since very little available software could be used without modification. The aim of this paper is to describe the new algorithms that were developed, to determine the convergence properties of our Monte Carlo integration method (Kirian *et al.*, 2010) and to evaluate the quality of one set of LCLS crystallographic data using metrics familiar to the crystallographic community. In this way we hope to define criteria for evaluating data from this new form of crystallography.

2. Monte Carlo intensity-merging theory

Fig. 1 shows a simulated X-ray snapshot diffraction pattern from a nanocrystal of Photosystem I (PSI, space group $P6_3$, $a = b = 281$, $c = 165$ Å) at 1.8 keV, to indicate the idealized features of a typical nanocrystal diffraction pattern [see Kirian *et al.* (2010) for details]. This is a fully spatially coherent simulation of a randomly oriented parallelepiped crystal of 17 × 17 × 30 unit cells (~0.5 μm in size), with a spectral width of 0.1% and beam divergence of 1.5 mrad. The diffracted intensity is given by the intersection of the Ewald sphere with the Fourier transform of the entire crystal, and in this case the crystals are small enough to produce a series of subsidiary intensity maxima surrounding each reciprocal-lattice point. While a single diffraction pattern contains only partially integrated reflection intensities, we describe below how structure factors may be extracted through averaging the intensities from many such partial reflections collected from crystals which vary in size, shape and orientation. Previous simulations (Kirian *et al.*, 2010) showed that, for a 10% size variation, this procedure can produce highly accurate structure factors with as few as some tens of thousands of nanocrystals in the absence of beam divergence, spectral width and crystal mosaicity. We term this Monte Carlo merging because we assume that, in the absence of a goniometer, a sufficient number of randomly oriented crystallites will sample all possible crystal orientations, sizes and shapes approximately uniformly. This assumption is supported by our previous simulations (Kirian *et al.*, 2010). Non-uniform orientation distributions are discussed in §7.

If we assume monochromatic plane-wave radiation with incident wavevector $\mathbf{k}_i$ ($|\mathbf{k}_i| = 1/\lambda$), the diffracted photon flux I_n (photons per pulse per pixel) at scattering vector $\Delta\mathbf{k} = \mathbf{k}_i - \mathbf{k}_o$

from the nth randomly oriented finite crystal is given in the kinematic theory as

$$I_n(\Delta\mathbf{k}) = J_o r_e^2 P |F(\Delta\mathbf{k})|^2 |S_n(\Delta\mathbf{k})|^2 \Delta\Omega, \tag{1}$$

where $F(\Delta\mathbf{k})$ is the transform of the average unit cell, $S_n(\Delta\mathbf{k})$ is the transform of the truncated crystal lattice (an interference function, similar to a three-dimensional sinc function laid down at reciprocal-lattice points), J_o is the average incident photon flux density (photons per pulse per area), r_e is the classical radius of the electron, P is a polarization factor and $\Delta\Omega$ is the solid angle subtended by the detector pixel. The vector $\Delta\mathbf{k}$ is defined in the crystal reference frame. We assume that absorption effects are negligible for our crystals of micron dimensions, although an absorption correction may be needed for the surrounding liquid in which the crystals are suspended.

For a given detector pixel, the observed $\Delta\mathbf{k}$ may be determined from the geometry of the detector and the crystal orientation. The average molecular transform $F(\Delta\mathbf{k})$ is defined here to be identical for all of the nanocrystals, but the lattice transform $S_n(\Delta\mathbf{k})$ depends on the size and shape of the crystal, and may differ significantly from one crystal to the next. However, we assume that the lattice transform always obeys the translational symmetry $S_n(\Delta\mathbf{k}) = S_n(\Delta\mathbf{k} + \mathbf{g}_{hkl})$, where $\mathbf{g}_{hkl}$ is any reciprocal-lattice vector with Miller indices hkl. For a perfect crystal, $S_n(\mathbf{g}_{hkl})$ is equal to the number of unit cells in the nth crystal, and $I_n(\mathbf{g}_{hkl})$ is therefore proportional to the square of the number of unit cells. The integrated lattice transform is proportional to the square root of the number of unit cells, and the integrated reflection intensity is proportional to the number of unit cells (Holton & Frankel, 2010).

The structure factors F_{hkl} we would like to extract from diffraction data are ideally equal to the unit-cell transform evaluated at a reciprocal-lattice point $\mathbf{g}_{hkl}$. Since the probability of observing diffraction precisely at $\mathbf{g}_{hkl}$ is essentially zero, we instead average intensities that fall within a small sphere centered at the lattice point. We define an integration domain radius δ, such that all intensities for which $|\Delta\mathbf{k} - \mathbf{g}_{hkl}| < \delta$ will be included in the average. After merging the indexed diffraction data in the three-dimensional diffraction volume from crystals that differ in size, shape and orientation, we may then form the average intensity

$$\langle I_n(\Delta\mathbf{k})\rangle_{n,\delta,hkl} = J_o r_e^2 P \langle |F(\Delta\mathbf{k})|^2 |S_n(\Delta\mathbf{k})|^2 \rangle_{n,\delta,hkl} \Delta\Omega, \tag{2}$$

where $\langle\rangle_{n,\delta,hkl}$ is understood to mean that we average over the distribution of crystal shapes and sizes, but only include intensity measurements for which $|\Delta\mathbf{k} - \mathbf{g}_{hkl}| < \delta$ is satisfied. If δ is sufficiently small we may write the approximation

$$\langle I_n(\Delta\mathbf{k})\rangle_{n,\delta,hkl} \simeq J_o r_e^2 P |F_{hkl}|^2 \langle |S_n(\Delta\mathbf{k})|^2 \rangle_{n,\delta,hkl} \Delta\Omega \tag{3}$$

and the structure-factor magnitudes may be evaluated as

$$|F_{hkl}|^2 \simeq \frac{\langle I_n(\Delta\mathbf{k})\rangle_{n,\delta,hkl}}{J_o r_e^2 P \langle |S_n(\Delta\mathbf{k})|^2 \rangle_{n,\delta} \Delta\Omega}. \tag{4}$$

We assume a well defined distribution of crystal shapes and size, so that there exists a mean value of $\langle |S_n(\Delta\mathbf{k})|^2 \rangle_{n,\delta}$. Since any lattice transform is identical when translated by a reciprocal-lattice vector $\mathbf{g}_{hkl}$, $\langle |S_n(\Delta\mathbf{k})|^2 \rangle_{n,\delta}$ is a constant that does not depend on the specific Miller indices hkl – an identical shape transform is laid down around every reciprocal-lattice point. We may therefore extract a quantity proportional to structure factors without any knowledge of the crystal size and shape distribution since

$$|F_{hkl}|^2 \propto \frac{\langle I_n(\Delta\mathbf{k})\rangle_{n,\delta,hkl}}{P\Delta\Omega}. \tag{5}$$

Accurate results can be expected from equation (5), provided that a reasonable integration domain radius δ is chosen, and that we measure a sufficient number of diffraction patterns to sample the various crystal shapes, sizes and orientations. The value of δ should be chosen to be smaller than the features of the unit-cell transform, which, according to the Shannon sampling theorem, corresponds to approximately $\delta < 1/2d$ for the largest cell constant d. Since intensities are averaged in a Monte Carlo fashion, relying on chance to provide all needed crystal orientations, shapes and sizes, the error in a measured mean intensity I is the standard error of the mean $\sigma(I)/(N-1)^{1/2}$, where N is the number of measurements (pixels) contributing to the particular intensity, and $\sigma(I)$ is the standard deviation in the intensity. An exceedingly small δ will drive down the value of N, while an oversized δ may increase the variance in intensities or sample unwanted background counts. The distribution of crystal sizes will have a particularly significant effect on the value of $\sigma(I)$, and a narrow size distribution is clearly preferable, unless the data are scaled according to crystal size prior to merging of intensities. Optimization of δ will depend on beam divergence, spectral width, crystal disorder, mosaicity and so on, all of which have been neglected in our simplified model above. Errors introduced during data processing should also be considered, as discussed in §7.

3. Description of the intensity-merging algorithm

In a previous paper (Kirian *et al.*, 2010) we described an algorithm for averaging structure-factor intensities, which we briefly review here. Each crystal orientation may be determined with autoindexing software, and is represented by the orientation matrix of the unit cell (Arndt & Wonacott, 1977),

$$\mathbf{A} = \begin{pmatrix} a_x^* & b_x^* & c_x^* \\ a_y^* & b_y^* & c_y^* \\ a_z^* & b_z^* & c_z^* \end{pmatrix}, \tag{6}$$

which specifies the crystal orientation (relative to the laboratory frame) through the reciprocal unit-cell vectors $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$. The reciprocal-lattice vector $\mathbf{g}_{hkl}$ may be related to $\mathbf{A}$ by the matrix product

$$\mathbf{g}_{hkl} = \mathbf{A}\mathbf{h}, \tag{7}$$

where the vectors $\mathbf{g}_{hkl}$ and $\mathbf{h}$ (which contains Miller indices) are understood to be column matrices. For a scattering vector

$\Delta\mathbf{k}_j$ (in the laboratory frame) corresponding to pixel j, fractional Miller indices $\mathbf{h}_j^{\rm frac}$ are determined by the equation

$$\mathbf{h}_j^{\rm frac} = \mathbf{A}^{-1}\Delta\mathbf{k}_j. \quad (8)$$

The nearest reciprocal-lattice vector to pixel j is taken to be $\mathbf{g}_j = \mathbf{A}\mathbf{h}_j$, where $\mathbf{h}_j$ are the integer values nearest to $\mathbf{h}_j^{\rm frac}$. The inequality $\delta > |\Delta\mathbf{k}_j - \mathbf{g}_{hkl}|$ may now be evaluated for each pixel in a detector, and the average intensity expressed in equation (5) may be calculated.

It is important to note that if $\mathbf{A}$ is determined from the locations of Bragg reflections alone, without consideration of diffracted intensities, multiple indexing convention choices may be possible. For space groups in which the Bravais lattice has higher symmetry than the point group, ambiguities in crystal orientations arise. If not properly accounted for, the apparent symmetry of the merged data will be higher than the true symmetry of the individual crystals. Specifically, in the present case the true space group is $P6_3$, but the merged data have apparent symmetry $P6_322$ and appear to have a 'twin fraction' of 0.5. (We assume that the nanocrystals themselves are not physically twinned; the apparent twinning is an artifact of indexing ambiguities in the data merging.) It is theoretically possible to 'de-twin' data from twinned crystals using this new data-collection method. Specifically, if nanocrystals do consist of single mosaic blocks then they cannot be twinned, even if the macroscopic crystals of the same form are merohedrally twinned. Once spot partialities are available, the histogram of 'estimated full' spot intensities will be bimodal for many *hkl* indices, allowing us to re-index each diffraction pattern to a consistent orientation choice, so that merging these data will produce an apparent symmetry of $P6_3$ and a twin fraction of 0. However, we have not attempted to perform this analysis here, and we instead treated these initial merged data as a 'twinned' data set.

4. Description of data

Data for this analysis were provided by the CFEL/ASG/ASU collaboration described elsewhere (Chapman *et al.*, 2011) using the CAMP instrument (Strüder *et al.*, 2010) on the AMO beamline (Bozek, 2009) at the LCLS, the first hard-X-ray FEL (Emma *et al.*, 2010). The nominal experimental parameters are briefly summarized. The X-ray pulse duration was 70 fs with more than 10^{12} photons per pulse incident on the nanocrystals (approximately the same number arriving in 1 s at a modern synchrotron). The average photon wavelength was 0.69 nm (1.8 keV energy) with a spectral width 0.1% FWHM. The X-ray beam was focused to a diameter of 7 µm, with a divergence of 0.5 mrad FWHM. Diffraction patterns were recorded on two fast-readout split-panel pnCCD detectors (Strüder *et al.*, 2010) with sample-to-detector distances of 67.8 mm to the wide-angle detector and 496 mm to a small-angle detector (only the wide-angle data were used here). Each detector consists of 512 × 1024 16-bit, 75 × 75 µm pixels lying in a plane normal to the beam, separated by a small gap through which the direct beam passed to a beam dump. The readout rate of the pnCCDs was matched to the 30 Hz pulse repetition rate of the FEL, resulting in over three million patterns after 28 h of operation. The resolution range (*d* spacing) for the wide-angle detector was 8.7–0.88 nm.

Full details of the Photosystem I nanocrystal sample preparation are given elsewhere (Chapman *et al.*, 2011; Hunter *et al.*, 2011). Briefly, fully hydrated PSI nanocrystals [Protein Data Bank (PDB) data entry 1JB0], within their native harvesting buffer [5 m*M* MES, where MES is 2-(*N*-morpholino)ethanesulfonic acid, pH 6.4 and 0.02% beta-dodecylmaltoside], were delivered to the X-ray beam *via* a 4 µm continuous column of liquid generated by a gas-dynamic virtual nozzle described elsewhere (DePonte *et al.*, 2008; Weierstall *et al.*, 2008). Elastic X-ray scattering from the liquid jet is concentrated into a streak running normal to the velocity of the liquid. A flow rate of 11 µl min^{-1} delivered 10 pmol of protein per minute at a concentration of 1 µ*M*, corresponding to 2.5 × 10^9 crystals ml^{-1}. A 2 µm in-line filter limits the maximum particle size that can pass through the nozzle in order to allow days of continuous jetting without clogging. Larger crystals are needle shaped, with approximately 5:1 aspect ratio. On average, crystal diffraction patterns were recorded at a hit rate of 110 per minute, where 'hits' are defined as patterns that contain ten or more detectable Bragg reflections. Settling of the crystals within the liquid delivery line causes the hit rate to drop exponentially with time (the theory of Brownian motion under an external force, described by the Smoluchowski equation, gives the settling time inversely proportional to the fifth power of the particle size). Hit rates may be adjusted by changing the nanocrystal concentration, although highly concentrated solution was avoided since patterns containing diffraction from multiple crystals are rejected by our current analysis method. Photosystem I nanocrystals crystallize in space group $P6_3$ with unit-cell lengths of nominally $a = b = 284.2$ and $c = 165.7$ Å, which lie between the parameters determined by single-crystal work at liquid-nitrogen temerature ($a = b = 281$ and $c = 165.2$ Å) (Jordan *et al.*, 2001) and 277 K ($a = b = 288$ and $c = 167$ Å) (Krauss *et al.*, 1996). Evaporative cooling in vacuum along the liquid jet may produce a temperature change between 10 and 100 K along the 200 µm length of the 5 µm-diameter liquid column. This drop in temperature may account for the observed unit-cell dimensions.

5. Data processing

All patterns were processed by removal of known bad pixels, applying a pre-characterized gain (flat-field) correction and subtracting a background. Typical combined background and readout signal was 0.9 photon equivalents per pixel, with a standard deviation of less than 1.3 photon equivalents. Background was estimated using a moving-window selection of 50 frames to ensure similar jet and exposure conditions. For each pixel, the background was taken as the median of values at that pixel over the 50 frames, which is insensitive to the relatively rare Bragg-diffracted photons. Because the jet streak occasionally drifted owing to motion of the liquid jet, its direction was determined by a linear fit to the streak, and a new jet mask was generated for each shot. Pixel saturation occurred at 80 photons per pixel. For the quantitative analyses of indexing and averaging intensities, saturated pixels were identified by the gross changes in peak shape which saturation produces, using pattern-recognition software. Saturated pixels in the pnCCD detectors tend to leak charge into neighboring pixels, so this was integrated and condensed into a small artificial peak at the centroid of the bloomed region. This step essentially increased the dynamic range of the detector, but shape information on the saturated spots was lost.

Following all of the pre-processing steps described above, each pattern was searched for Bragg reflection peaks by threshold analysis and aperture photometry. Convolution of each image with a top-hat function provided a locally integrated intensity, while convolution with a concentric annulus provided an estimate of local background that was subtracted. Bragg peaks were identified by applying a threshold of 25 to the resulting locally integrated photons, and positions were taken to be the centroid of nearby intensities. Lists of peak coordinates were processed by the program *MOSFLM* (Leslie, 1992, 2006) using the *DPS* autoindexing algorithm (Steller *et al.*, 1997), which resulted in an estimated **A** matrix for each crystal. Each **A** matrix was further refined by global optimization of the unit-cell constants and three Euler angles, which was carried out by minimization of r.m.s. residual distances between predicted and observed reciprocal-space peak positions using a simplex algorithm. Photon wavelengths were estimated shot by shot based on the measured electron-beam energy, current and the undulator K value of the LCLS. The orientation and position of each of the two detector panels were initially determined from powder patterns constructed by summation of many snapshot images from lysozyme nanocrystals. These gave higher accuracy than the PSI data owing to larger spacing between adjacent reflections. Further refinement of the detector geometry was performed on a shot-by-shot basis by constrained global optimization of the detector coordinates, similar to the method applied to the refinement of the **A** matrix.

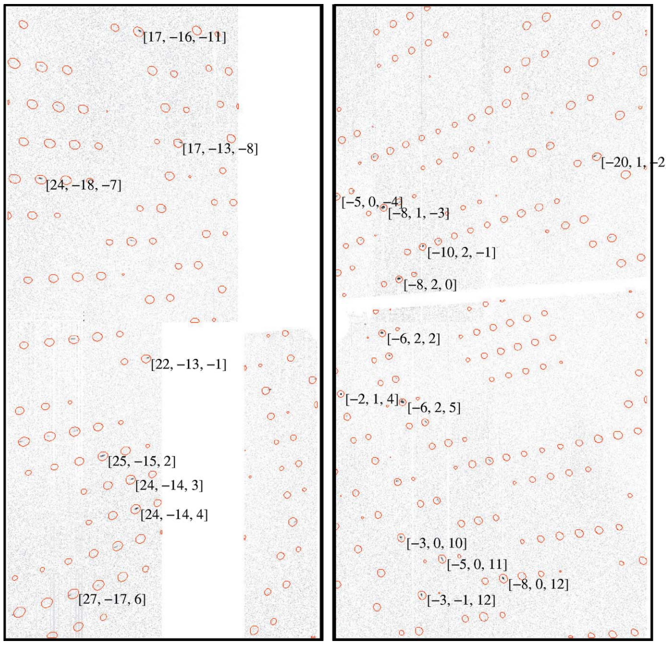


Figure 2
Typical processed diffraction pattern showing both of the near pnCCD panels on a logarithmic intensity scale. All rejected pixels (including those near the streak caused by the liquid jet) have been masked. Red rings indicate integration domains for a threshold of δ = 8.7 μm^{-1}, after processing with *MOSFLM*, and additional refinement of cell constants and crystal orientation. Miller indices are indicated for a selection of bright Bragg reflections. There is a substantial gap (the vertical line down the middle) between the two panels, through which the direct beam passes.

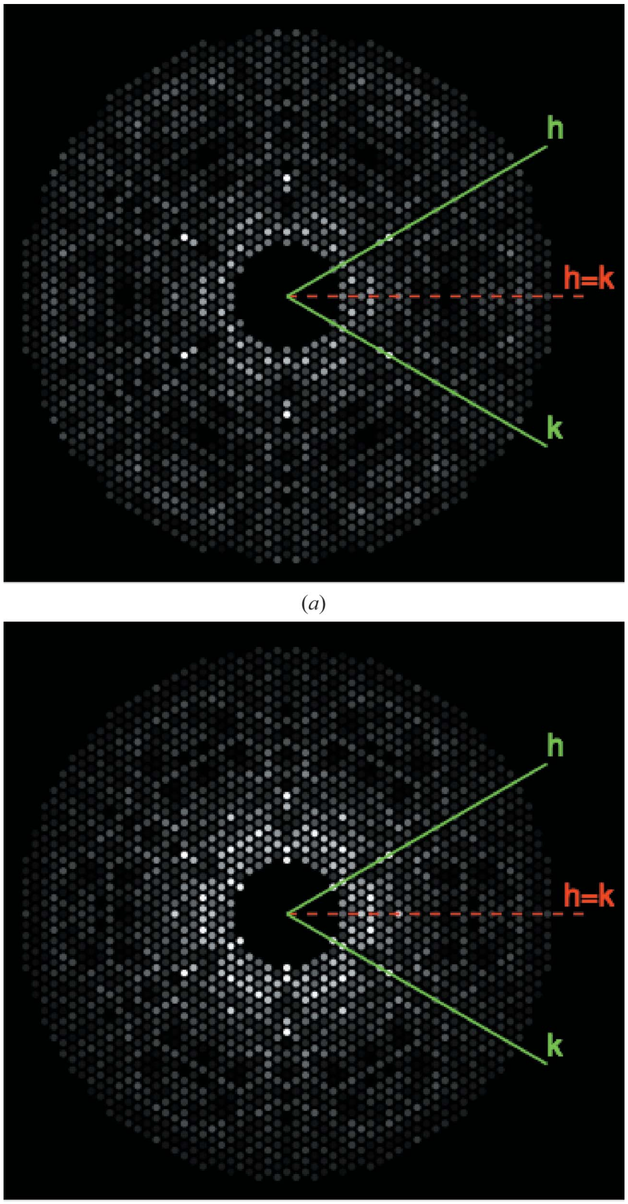
Structure factors were extracted according to the algorithm described in §3. Since the LCLS beam is strongly linearly polarized, a polarization factor of $P(\mathbf{k}_o) = 1 - |\hat{\mathbf{u}} \cdot \hat{\mathbf{k}}_o|^2$ was assumed, where the vector $\hat{\mathbf{u}}$ points along the horizontal direction and the vector $\hat{\mathbf{k}}_o$ points from the interaction region to a detector pixel. Pixel solid angles were approximated with the expression

$$\Delta\Omega = \left(\frac{l}{L}\right)^2 \mathbf{n} \cdot \hat{\mathbf{k}}_o, \tag{9}$$

where L is the distance from the sample to the pixel, l is the pixel size and $\mathbf{n}$ the unit vector normal to the detector plane. During intensity averaging, patterns with multiple crystals or poorly determined crystal orientation were rejected by requiring that 80% of the observed peaks must lie within a small distance (10 μm^{-1}) from predicted peak locations. All pixels that were marked as bad during pre-processing were excluded from the averaged intensities.

Of the 1.85 million diffraction patterns collected at 70 fs pulse duration, 112 725 patterns contained ten or more identifiable peaks. A total of 28 192 were indexed with unit-cell parameters within 5% of the mean values. Indexing success was strongly correlated with the number of detected peaks. Only 5% of patterns with ten peaks were successfully indexed, whereas more than 60% of patterns with more than 75 peaks were successfully indexed. Approximately 42% of indexed patterns were rejected during intensity merging because of disagreement between predicted and observed peak locations, resulting in a total of 16 374 patterns that contributed to the measured structure-factor amplitudes presented here. Data processing was carried out on a desktop computer with a 2.26 GHz quad core Intel Xeon processor. For the steps of jet masking, peak finding, indexing and intensity merging, the overall processing speed is typically less than 4 s per pattern when using a single processor core.

Fig. 2 shows a typical diffraction pattern after pre-processing, background subtraction, and masking of bad pixels and the jet streak. Smaller crystals tend to display more prominent shape transforms, similar to the simulation in Fig. 1. Red rings indicate the domains within which intensities contribute to the derived structure-factor amplitudes, according to equation (5), for an integration domain radius of δ = 8.7 μm^{-1}. A calculated precession image on the [001] zone axis is shown on a logarithmic scale in Fig. 3(*b*), for an integration domain radius $\delta = 7.3\ \mu\text{m}^{-1}$. Equivalent reflections of the $P6_322$ space group have been merged, which results in the perfect sixfold symmetry and mirror symmetry about the $h = k$ line (the 'apparent twinning' described in §3). In Fig. 3(*a*) the same subset of intensities is shown for single-crystal data obtained with 1.0 Å wavelength X-rays at the Advanced Light Source (ALS) in Berkeley (Fromme & Fromme, 2010), which have been artificially 'twinned' for comparison with the LCLS data by taking the mean intensity values of the hkl and $kh\bar{l}$ reflections. We discuss and evaluate differences between these data sets in the next section.



(*a*)

(*b*)

Figure 3

(*a*) Calculated precession image of the [001] zone extracted from single-crystal data obtained at the Advanced Light Source. 50% twinning has been applied. (*b*) Calculated precession image of the [001] zone extracted from PSI nanocrystal data collected at the LCLS following indexing by *MOSFLM*. The data set consists of 16 374 single snapshot diffraction patterns of 70 fs pulse length collected at the AMO beamline at LCLS at 6.9 Å wavelength. The space group was $P6_3$, but $P6_322$ symmetry was imposed by autoindexing ambiguity, resulting in an apparent twin fraction of 0.5 for this data set. The integration domain radius was $\delta = 7.3\ \mu\text{m}^{-1}$.

6. Quality of data, Monte Carlo convergence

In order to assess the internal consistency of our structure factors, we have computed a figure-of-merit $R_{\rm int}$, which compares structure factors $F_{\rm even}$ extracted from the even-numbered diffraction patterns with the odd-numbered $F_{\rm odd}$. The quantity $R_{\rm int}$ is defined as

$$R_{\rm int} = \frac{\sum_{\{hkl\}} \big| |F_{\rm even}| - |F_{\rm odd}| \big|}{\sum_{\{hkl\}} \big| (|F_{\rm even}| + |F_{\rm odd}|)/2 \big|}. \qquad (10)$$

Structure factors were taken to be the square root of intensities obtained through equation (5). Fig. 4 shows $R_{\rm int}$ as a function of ordinal image number for several values of the intensity integration volume defined by δ. It falls smoothly to a value of less than 5% for the largest integration domain. The decrease in $R_{\rm int}$ corresponding to increasing values of δ may be due, in part, to increasing numbers of background counts in the integration (see §7 for further discussion).

We have also analyzed the variation of $R_{\rm int}$ with resolution. This is shown in Fig. 5, together with a correlation coefficient $C_{\rm int}$ defined by

$$C_{\rm int} = \frac{\sum_{\{hkl\}} \big(F_{\rm even} - \langle F_{\rm even}\rangle\big)\big(F_{\rm odd} - \langle F_{\rm odd}\rangle\big)}{\sum_{\{hkl\}} \Big[\big(F_{\rm even} - \langle F_{\rm even}\rangle\big)^2\Big]^{1/2} \Big[\big(F_{\rm odd} - \langle F_{\rm odd}\rangle\big)^2\Big]^{1/2}}. \qquad (11)$$

The rise in $R_{\rm int}$ at about 9.4 Å resolution is due to our use of a square detector, offering greater resolution in diagonal directions, but fewer recorded reflections at the corners of the detector. Prior to this, the $R_{\rm int}$ appears to converge to values

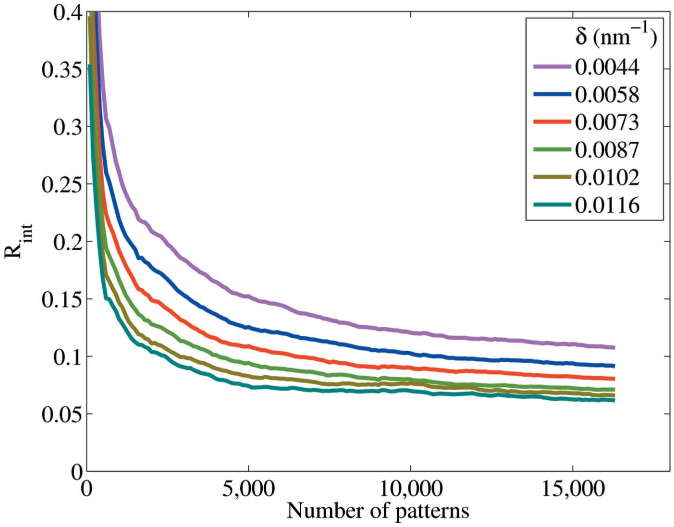


Figure 4

Consistency between merged structure factors obtained from even- and odd-numbered image frames, plotted as a function of ordinal image number. $R_{\rm int}$ is shown for various integration domain radii δ.

between 6 and 12%, suggesting that high-resolution data lie beyond the range of our current detector (which cuts off at about 8.8 Å). Errors in merged intensities were assessed by calculating the r.m.s. difference $\sigma(I)$ between even- and odd-frame reflection intensities. The median value of $I/\sigma(I)$ within 9–11 Å resolution is 10.3, well above the customary cutoff value of about 2.

Since the above internal consistency measures do not reveal information about the *accuracy* of extracted structure factors, we have conducted rigid-body refinement of the published PSI structure (1JB0) against the LCLS structure factors using *REFMAC* (Murshudov *et al.*, 1997). This refinement was conducted in 'twin mode' and with bulk solvent contribution (Tronrud, 1997; Afonine *et al.*, 2005), and the resulting *R*/*R*-free values were 0.284/0.327. This procedure fitted a total of 84 parameters (overall scale and anisotropic *B* factor, the bulk solvent scale and *B* factor, 12 rigid protein chains with three translations and three rotations each, and one twin fraction) to 6248 observations. By comparison, the *R*/*R*-free from refining 1JB0 against data obtained from a cryo-cooled single crystal of PSI at ALS were 0.285/0.298. The similarity of these refinement results suggests that the structure factors we obtained at LCLS are at least as consistent with the native structure of PSI as conventional single-crystal data to the resolution available (9 Å). Unfortunately, we do expect considerable non-isomorphism between the crystals used at ALS and LCLS because the solution conditions and data-collection temperatures were different (Crick & Magdoff, 1956; Blundell & Johnson, 1976). Indeed, the unit-cell dimensions changed by 1% and the overall $R_{\rm iso}$ between these two (LCLS, ALS) data sets was 23.5% (see also Fig. 5 in which we define $R_{\rm iso}$ and $C_{\rm iso}$ in the same way as $R_{\rm int}$ and $C_{\rm int}$, but with even/odd structure factors replaced with those obtained at LCLS/ALS). The difference is very likely caused by two effects: the cryoprotectant and the cooling process itself. The LCLS data were collected from nanocrystals in their low-ionic-strength mother liquor in the non-frozen state, while the ALS data were collected from crystals in 2 *M* sucrose low-ionic-strength buffer at cryogenic temperatures (96 K). As the crystals contain 78% solvent the replacement of the mother liquor with 2 *M* sucrose in ten steps over a time period of 2 h prior to flash freezing in liquid propane may contribute as much to the slight non-isomorphism as the cooling process. As shown in Fig. 6, plots of the asymptotic $R_{\rm iso}$ value (after merging 16 374 patterns) against δ show a minimum value for $\delta = 7.3\ \mu{\rm m}^{-1}$.

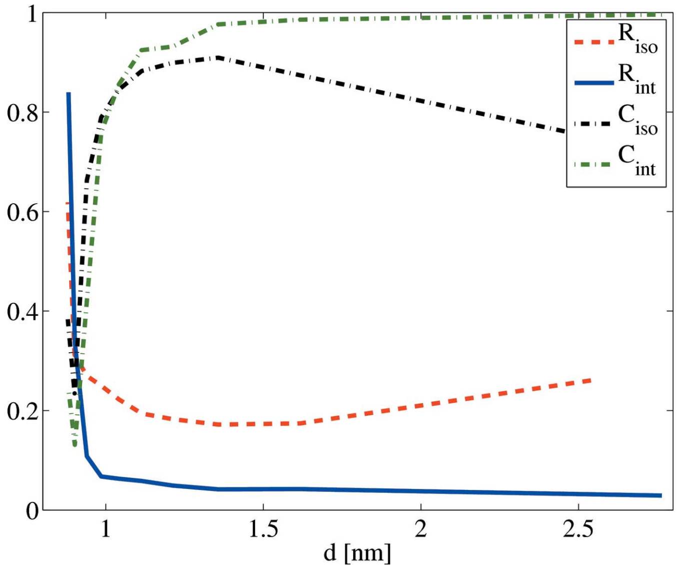


Figure 5
R factors and linear correlations *C* (for integration domain $\delta = 7.3\ \mu{\rm m}^{-1}$) plotted against resolution. Approximately 350 structure factors fall within each resolution bin. The subscript 'iso' refers to comparison of conventional single-crystal diffraction data to fs nanocrystal diffraction data. The subscript 'int' refers to comparison of even to odd frames from fs nanocrystal diffraction data.

We note that the absorbed dose per PSI crystal for a single 70 fs shot is significantly larger (about 405 MGy) than the recommended safe dose limit of 30 MGy for cryo-cooled protein crystals at a third-generation light source (Owen *et al.*, 2006). The initial analysis of the PSI nanocrystal diffraction data described recently by Chapman *et al.* (2011) demonstrates an apparent lack of radiation damage to a detector-limited resolution of 0.9 nm when using 70 fs pulses or shorter, despite the subsequent destruction of the nanocrystals by the photoelectron cascade following termination of the XFEL pulse (Chapman *et al.*, 2006).

Data completeness is demonstrated in Fig. 7, as a function of the number of patterns and δ. Completeness is measured in the $P6_322$ space group since the two possible indexing orientations cannot be distinguished (see §3). For our highly symmetric space group (of the 70 082 total reflections, only 3328 are unique), the data set was complete using less than 10 000 patterns, depending on the δ value.

Fig. 8 shows a density map for PSI constructed using the program *REFMAC*. Rigid-body displacements were applied to the PSI structure 1JB0 in the PDB for best fit to the LCLS set of structure factors to obtain this map.

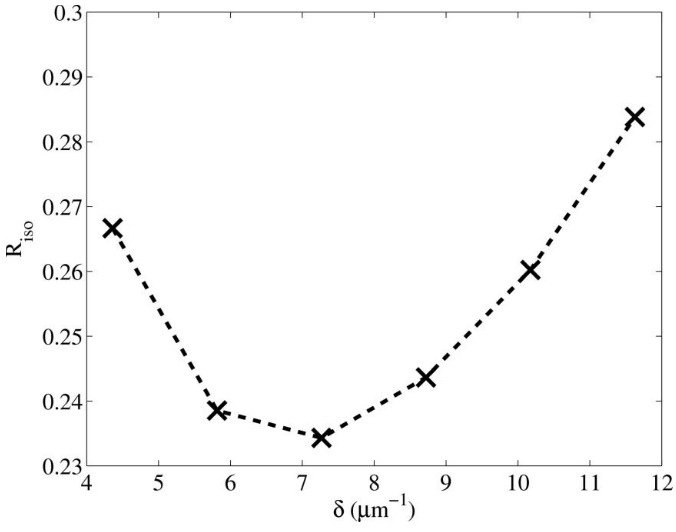


Figure 6
Asymptotic value of $R_{\rm iso}$, comparing LCLS data with single-crystal data, plotted against the integration volume δ used around each reciprocal-lattice point for these nanocrystals of varying size. Despite the fixed integration volume but varying crystal size, a minimum is found.

7. Discussion

The effects of crystal orientation and refinement of the Euler angles discussed following equation (1) may need further consideration in future analysis of structure-factor extraction from nanocrystal data. In comparison to data from large single crystals, the structure factors determined here tend to have a more rapid decay in intensity with scattering angle, which is not apparent when merging the data into a virtual powder pattern. The most likely cause of this decay is the effective increase in integration domain at higher scattering angles due to orientation errors that allow high-angle spots to 'escape' from the integration region more often than low-angle spots. This problem is particularly severe for patterns that have only a very limited number of reflections. Orientation errors may be assessed by merging all diffraction patterns into a single, oversampled three-dimensional intensity map, followed by a measurement of the average peak broadening of Bragg reflections as a function of resolution. Although we have not carried out such an analysis here, applying artificial scale factors of the form $\exp(-B|\Delta\mathbf{k}|)$ (similar to a temperature factor 'correction') suggests that it may be possible to decrease R and C somewhat if these errors can be corrected. However, these problems may be alleviated in future work by increasing the number and quality of peaks used in indexing [*e.g.* by increasing photon energy, pulse intensity, crystal size, detector dynamic range, and improving the peak-finding and peak-indexing algorithms (Maia *et al.*, 2010)]. With increasingly accurate crystal orientations, the integration domain may be decreased significantly, which will likely result in further improvements in the accuracy of merged intensities.

Peak widths, as determined from virtual powder patterns, are approximately 2.68×10^{-4} Å^{-1} FWHM, giving a minimum nanocrystal size of approximately 0.372 µm, which is consistent with the smaller crystal size estimates based on the interference fringes between Bragg peaks. The instrumental effects of spectral broadening and beam divergence, 0.21 and 0.5 mrad FWHM, respectively, are therefore negligible for the present analysis. We assume here that each nanocrystal is smaller than one mosaic block (Hunter *et al.*, 2011). We have thus not attempted to correct for any effects of crystal mosaicity in the present analysis, but we note that if crystals consist of multiple mosaic domains, then the effect of mosaicity on the merged intensities will be similar in nature to that of crystal-orientation errors, and the combined effect of both may perhaps be corrected for simultaneously in future analysis.

Although we have not observed significant signs of flow alignment of PSI crystals caused by the liquid jet, we anticipate that these effects may need to be considered during data analysis for certain protein crystals. Since structure factors are derived from averaged diffracted intensities (each integrated reflection intensity is divided by the number of observations), our method has no effective Lorentz factor correction and is likely to be insensitive to the degree of flow alignment. In addition, optical examination of the smaller crystallites shows them to be highly irregular in shape, faceted and with an aspect ratio tending to unity. While strongly flow-aligned data will result in a reduced completeness because of the blind regions at either pole of reciprocal space (much like the blind regions in typical oscillation data), this problem may be easily remedied by adjusting the angle of the liquid jet relative to the X-ray beam. However, if the degree of alignment is correlated with crystal size (*e.g.* if larger crystals tend to have a larger aspect ratio resulting in a greater degree of alignment), the effects of flow alignment will need to be considered more carefully, perhaps by first classifying crystals by size (*e.g.* using the high angular resolution shape transform on the small-

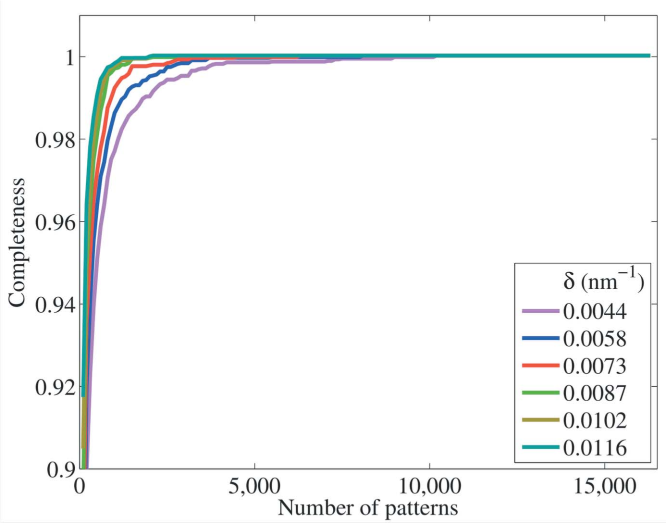


Figure 7
Completeness of the data as a function of the number of patterns shown for several values of the integration parameter δ. Note that the completeness is measured as the fraction of unique reflections which have been observed at least once, divided by the total number of unique reflections within the resolution range of the detector. Unique reflections means $P6_322$, which is highly symmetric. Lower-symmetry space groups will require far more patterns in order to collect a complete data set.

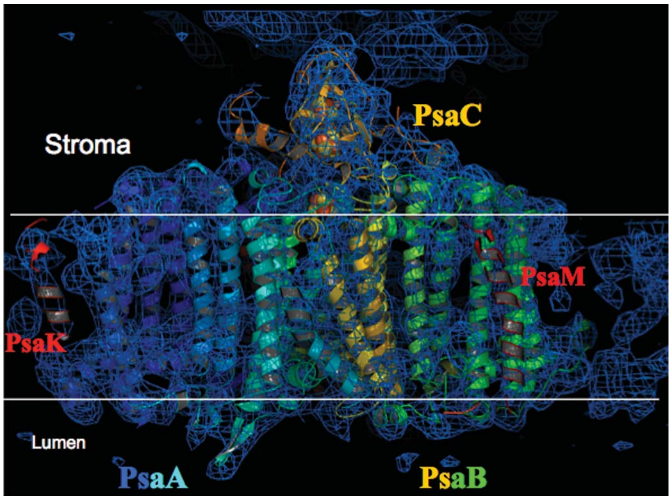


Figure 8
Density map for PSI obtained by combining this LCLS nanocrystal diffraction data with data in the PDB. Some of the 12 proteins in each monomer are labelled PsaA *etc*. The position of the membranes is indicated by lines across the figure.

angle detector), carrying out separate analysis for each size class and scaling each data set appropriately prior to merging. The degree of flow alignment in a data set may be determined straightforwardly from autoindexing results. Even in the absence of flow alignment, scaling of multiple data sets according to crystal size classes should be investigated, as it will likely increase the rate of convergence and the accuracy of the resulting merged intensities.

We note that the rate of convergence for the experimental data is much greater than for our simulated data (Kirian *et al.*, 2010), despite a higher final R factor. This may result from the absence of beam divergence, crystal disorder and beam energy spread in our simulations.

8. Conclusions

The aim of this work has been to study the convergence of our Monte Carlo integration method using a limited subset of the data collected at LCLS in December 2009. Since both $R_{\rm int}$ and $C_{\rm int}$ apparently tend to an asymptote, there may be little reason to collect more than about 10 000 diffraction patterns of Photosystem I crystals for a complete data set to a resolution of 0.9 nm. (Many nanocrystals are, however, wasted during flow in the period between X-ray pulses.) While a complete data set can, in principle, be collected in less than 6 min at the present 30 Hz repetition rate, the throughput in the present work is lower, owing to the necessity to target a hit rate of 20%, so that patterns with multiple crystals are minimized. A larger number of patterns will likely be needed for data collected at higher resolution, but this number depends on the intensity and wavelength of the X-ray beam, the crystal-size distribution and numerous other factors. However the power of the method may lie in its ability to take advantage of the wide range of microscopic crystals of various sizes which exist in the mother liquor used for crystal growth, or the showers of microcrystals sometimes seen during growth. For difficult-to-crystallize proteins, such as membrane proteins, it may take decades from the observation of the first showers of microcrystals to structure determination at atomic resolution. In the case of Photosystem I, it took 13 years from the first observation of microcrystals in 1988 (Witt *et al.*, 1988) to the structure determination at 2.5 Å resolution (Jordan *et al.*, 2001). The new avenue of fs nanocrystallography, where the structure determination is based on a fully hydrated stream of nanocrystals, may thereby improve the efficiency of protein structure analysis. In addition, by supplying the jet from a high-performance liquid chromatography (HPLC) apparatus and an autosampler, convenient control of the chemical environment is achieved for samples near room temperature and the method is directly applicable for high-throughput data collection, without the need to mount or freeze individual crystals. Because radiation-damage effects are avoided when using sufficiently short pulses, the need for cooling in protein crystallography is eliminated. The continuous replenishment of nanocrystals *via* the liquid nozzle also makes possible time-resolved studies of irreversible processes by extending the experimental apparatus to include *e.g.* an optical pump laser, or a mixing cell in which multiple reactants are mixed immediately upstream of the nozzle exit. Finally, preliminary indications are that, for short pulses, a dose much larger than the safe dose normally tolerable in protein crystallography is possible.

We are particularly grateful for a careful reading of the manuscript and corrections by Professors I. Schlichting, J. Ulrich and J. Hajdu. We acknowledge support from the National Science Foundation (NSF) CBST center at UC Davis, DOE award No. DE-SC0002141, NSF award No. 0417142, the US Department of Energy by Lawrence Livermore National Laboratory under contract No. DE-AC52-07NA27344, the Joachim Herz Stiftung, the Max Planck Society, the Swedish Research Councils, the Swedish Foundation for International Cooperation in Research and Higher Education, Stiftelsen Olle Engkvist Byggmästare, the Helmholtz Association (VH-VI-302) and the DFG Cluster of Excellence at the Munich Centre for Advanced Photonics. The Advanced Light Source is supported under contract No. DE-AC02-05CH11231 and ALS 8.3.1 by the National Institutes of Health (GM074929 and GM082250). Portions of this research were carried out at the Linac Coherent Light Source, a national user facility operated by Stanford University on behalf of the US Department of Energy, Office of Basic Energy Sciences.

References

Afonine, P. V., Grosse-Kunstleve, R. W. & Adams, P. D. (2005). *Acta Cryst.* D**61**, 850–855.

Arndt, U. W. & Wonacott, A. J. (1977). Editors. *The Rotation Method in Crystallography: Data Collection from Macromolecular Crystals.* Amsterdam: North-Holland.

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography.* New York: Academic Press.

Bozek, J. D. (2009). *EPJ ST*, **169**, 129–132.

Chapman, H. N. (2009). *Nat. Mater.* **8**, 299–301.

Chapman, H. N. *et al.* (2006). *Nat. Phys.* **2**, 839–843.

Chapman, H. N. *et al.* (2011). *Nature* (*London*), **470**, 73–77.

Crick, F. H. C. & Magdoff, B. S. (1956). *Acta Cryst.* **9**, 901–908.

DePonte, D. P., Weierstall, U., Schmidt, K., Warner, J., Starodub, D., Spence, J. C. H. & Doak, R. B. (2008). *J. Phys. D*, **41**, 195505.

Emma, P. *et al.* (2010). *Nat. Photonics*, **4**, 641–647.

Fromme, P. & Fromme, R. (2010). Unpublished.

Holton, J. M. & Frankel, K. A. (2010). *Acta Cryst.* D**66**, 393–408.

Howells, M. R., Beetz, T., Chapman, H. N., Cui, C., Holton, J. M., Jacobsen, C. J., Kirz, J., Lima, E., Marchesini, S., Miao, H., Sayre, D., Shapiro, D. A., Spence, J. H. C. & Starodub, D. (2009). *J. Electron Spectrosc. Relat. Phenom.* **170**, 4–12.

Hunter, M. S., DePonte, D. P., Shapiro, D. A., Kirian, R. A., Wang, X., Starodub, D., Marchesini, S., Weierstall, U., Doak, R. B., Spence, J. C. H. & Fromme, P. (2011). *Biophys. J.* **100**, 198–206.

Jordan, P., Fromme, P., Witt, H. T., Klukas, O., Saenger, W. & Krauss, N. (2001). *Nature* (*London*), **411**, 909–917.

Kirian, R. A., Wang, X. Y., Weierstall, U., Schmidt, K. E., Spence, J. C. H., Hunter, M., Fromme, P., White, T., Chapman, H. N. & Holton, J. (2010). *Opt. Express*, **18**, 5713–5723.

Krauss, N., Schubert, W. D., Klukas, O., Fromme, P., Witt, H. T. & Saenger, W. (1996). *Nat. Struct. Biol.* **3**, 965–973.

Laue, M. von (1936). *Ann. Phys.* **26**, 55–68.

Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EACMB Newsl. Protein Crystallogr.* **26**.

Leslie, A. G. W. (2006). *Acta Cryst.* D**62**, 48–57.

Maia, F. R. N. C., Yang, C. & Marchesini, S. (2010). *Ultramicroscopy*, doi:10.1016/j.ultramic.2010.10.016.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. (2000). *Nature* (*London*), **406**, 752–757.

Owen, R. L., Rudino-Pinera, E. & Garman, E. F. (2006). *Proc. Natl Acad. Sci. USA*, **103**, 4912–4917.

Rossmann, M. G. & Erickson, J. W. (1983). *J. Appl. Cryst.* **16**, 629–636.

Rossmann, M. G., Leslie, A. G. W., Abdel-Meguid, S. S. & Tsukihara, T. (1979). *J. Appl. Cryst.* **12**, 570–581.

Solem, J. C. (1986). *J. Opt. Soc. Am. B*, **3**, 1551–1565.

Spence, J. C. H. & Doak, R. B. (2004). *Phys. Rev. Lett.* **92**, 198102.

Steller, I., Bolotovsky, R. & Rossmann, M. G. (1997). *J. Appl. Cryst.* **30**, 1036–1040.

Strüder, L. *et al.* (2010). *Nucl. Instrum. Methods Phys. Res. Sect. A*, **614**, 483–496.

Tronrud, D. E. (1997). *Methods Enzymol.* **277**, 306–319.

Weierstall, U., Doak, R. B., Spence, J. C. H., Starodub, D., Shapiro, D., Kennedy, P., Warner, J., Hembree, G. G., Fromme, P. & Chapman, H. N. (2008). *Exp. Fluids*, **44**, 675–689.

Witt, I., Witt, H. T., Difiore, D., Rogner, M., Hinrichs, W., Saenger, W., Granzin, J., Betzel, C. & Dauter, Z. (1988). *Ber. Bunsenges. Chem. Chem. Phys.* **92**, 1503–1506.

Woolfson, M. M. (1997). *An Introduction to X-ray Crystallography.* Cambridge University Press.