Acta Crystallographica Section A
Foundations and Advances
ISSN 2053-2733

Received 20 May 2013
Accepted 2 December 2013

Direct phasing in femtosecond nanocrystallography. II. Phase retrieval

Joe P. J. Chen,[a] John C. H. Spence[b] and Rick P. Millane[a]*

[a]Computational Imaging Group, Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand, and [b]Department of Physics, Arizona State University, Tempe, AZ 85287, USA. Correspondence e-mail: rick.millane@canterbury.ac.nz

X-ray free-electron laser diffraction patterns from protein nanocrystals provide information on the diffracted amplitudes between the Bragg reflections, offering the possibility of direct phase retrieval without the use of ancillary experimental diffraction data [Spence *et al.* (2011). *Opt. Express*, **19**, 2866–2873]. The estimated continuous transform is highly noisy however [Chen *et al.* (2014). *Acta Cryst.* A**70**, 143–153]. This second of a series of two papers describes a data-selection strategy to ameliorate the effects of the high noise levels and the subsequent use of iterative phase-retrieval algorithms to reconstruct the electron density. Simulation results show that employing such a strategy increases the noise levels that can be tolerated.

1. Introduction

The three major problems that plague protein crystallography are crystal preparation, radiation damage and phase determination. X-ray free-electron lasers (XFELs) provide the potential to solve all three problems by the production of intense but extremely brief X-ray pulses. Appropriate signal levels can be attained whilst sidestepping the resolution-limiting effects of radiation damage, as the pulse is so brief that it terminates before significant radiation damage develops (Neutze *et al.*, 2000; Barty *et al.*, 2011). The high-intensity X-ray pulse enables measurable diffraction data to be obtained from nanocrystals only a few unit cells across.

Solution of the phase problem from such data still presents a problem however. Molecular replacement has recently been used to solve the structure of *Trypanosoma brucei* cysteine protease cathepsin B from XFEL data (Redecke *et al.*, 2013). It is not clear if the method of isomorphous replacement can be suitably adapted because of the experimental difficulty of obtaining isomorphous nanocrystals. Anomalous dispersion phasing of XFEL nanocrystal data that incorporates the effects of ionization damage of heavy atoms in the presence of the high-fluence XFEL pulse has been proposed (Son *et al.*, 2011), although its practical utility still requires experimental verification. Therefore, alternative methods of phasing are of significant interest.

The diffraction pattern from a nanocrystal has significant, but weak, amplitude between the Bragg peaks (Vartanyants & Robinson, 2001). Spence *et al.* (2011) have shown that the continuous molecular transform can be estimated from such data, which provides a possible route to direct reconstruction of the electron density using phase-retrieval algorithms (Miao & Sayre, 2000). In the first paper of this series (Chen *et al.*, 2014) we analysed in detail properties of diffraction patterns from a collection of nanocrystals of various sizes and their noise characteristics, and the consequences for estimating the molecular transform from the diffraction data. In this second paper we develop a strategy for applying iterative phase retrieval to the estimated molecular diffraction amplitudes at low signal-to-noise ratio (SNR) for structure determination. The methods described are evaluated by simulation and the results give a picture of the effects of SNR and mean nanocrystal size on reconstruction of the electron density.

2. Estimating the molecular transform

The first step in reconstructing the electron density is to estimate the molecular transform (the Fourier transform of the contents of one unit cell) from the nanocrystal diffraction data. The characteristics of this problem are described in detail by Chen *et al.* (2014) and are summarized briefly in this section. Here we consider only crystals consisting of an integral number of a single kind of unit cell, although recent studies indicate that the effect of different and incomplete unit cells, that can occur when there is more than one molecule in the unit cell, may be small (Chen & Millane, 2013; Liu *et al.*, 2014).

The intensity diffracted by a single three-dimensional crystal of size $N_1 \times N_2 \times N_3$ unit cells along the respective crystal axes, $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$, is given by (Spence *et al.*, 2011; Chen *et al.*, 2014)

$$I(\mathbf{N}, \mathbf{u}) = |F(\mathbf{u})|^2 S^2(\mathbf{N}, \mathbf{u}), \tag{1}$$

where $\mathbf{u}$ is the position vector in reciprocal space, $\mathbf{N} = (N_1, N_2, N_3)$ is a column vector of the number of unit cells in each crystal direction and $S(\mathbf{N}, \mathbf{u})$ is referred to as the shape transform function given by

$$S(\mathbf{N}, \mathbf{u}) = \prod_{i=1}^{3} \frac{\sin(\pi N_i \mathbf{a}_i \cdot \mathbf{u})}{\sin(\pi \mathbf{a}_i \cdot \mathbf{u})}. \quad (2)$$

In femtosecond nanocrystallography, typically tens of thousands of usable diffraction patterns are collected from crystals of varying sizes, shapes and orientations as they pass randomly in front of the pulsing XFEL beam. The orientation of each pattern can be determined *via* auto-indexing algorithms (Kirian *et al.*, 2010), and upon summing together patterns of the same orientation class the averaged diffracted intensity, denoted $\langle I_n(\mathbf{u})\rangle_n$, is given by (Chen *et al.*, 2014)

$$\left\langle I_n(\mathbf{u})\right\rangle_n = \left\langle I(\mathbf{N}_n, \mathbf{u})\right\rangle_n = |F(\mathbf{u})|^2 Q^2(\mathbf{u}), \quad (3)$$

where $\mathbf{N}_n$ is the crystal size for the nth nanocrystal, $\langle\cdot\rangle_n$ denotes averaging over the diffraction snapshots and the averaged shape transform $Q^2(\mathbf{u})$ is given by

$$Q^2(\mathbf{u}) = \sum_{\mathbf{N}} P(\mathbf{N}) S^2(\mathbf{N}, \mathbf{u}), \quad (4)$$

where $P(\mathbf{N}) = P(N_1, N_2, N_3)$ is the probability density function (p.d.f.) describing the distribution of crystallite sizes.

If the crystallite sizes in each of the three directions are independent, then equation (4) reduces to

$$Q^2(\mathbf{u}) = \prod_{i=1}^{3} \sum_{N_i} P(N_i) S^2(N_i, u_i). \quad (5)$$

In practice, the nanocrystals will not all be parallelopipeds but will adopt a variety of sizes and shapes, and will also be subject to various forms of disorder (Dilanian *et al.*, 2013; Chen *et al.*, 2014). However, the end result is that the averaged intensity is described by equation (3) where $Q^2(\mathbf{u})$ is appropriately modified to encapsulate all of these effects (Chen *et al.*, 2014).

Referring to equation (3), the magnitude of the molecular transform can be obtained from the data by

$$|F(\mathbf{u})|^2 = \frac{\left\langle I_n(\mathbf{u})\right\rangle_n}{Q^2(\mathbf{u})}. \quad (6)$$

The averaged shape transform $Q^2(\mathbf{u})$ cannot be calculated using equation (4) because the size distribution of the nanocrystals $P(\mathbf{N})$ is not known *a priori*. However, it can be shown that the averaged shape transform can be estimated directly from the diffraction data by averaging the diffracted intensities around each Bragg reflection, and is given by (Spence *et al.*, 2011; Chen *et al.*, 2014)

$$Q^2(\mathbf{u}) \propto \left\langle I_n(\mathbf{u} - \mathbf{u}_\mathrm{b})\right\rangle_{\mathrm{b},n}, \quad (7)$$

where $\mathbf{u}_\mathrm{b}$ is the position of the Bragg reflection and $\langle\cdot\rangle_{\mathrm{b},n}$ denotes averaging over the region around each Bragg reflection and over all diffraction snapshots. The estimate of the averaged shape transform from equation (7) can then be substituted into equation (6), giving an estimate of the molecular transform from the measured intensity alone. The noise characteristics of the diffraction and the derived molecular transform are described in detail by Chen *et al.* (2014) and are discussed further in §4.

3. Phase retrieval

Once an estimate of the continuous Fourier amplitude has been obtained, it is well known that in principle the electron density (or image) can be reconstructed by using real-space constraints, usually through the application of iterative projection algorithms (Sayre, 1952; Fienup, 1982; Bates & McDonnell, 1989; Millane, 1990; Miao *et al.*, 1999; Elser, 2003*a*,*b*; Martin *et al.*, 2012; Rodriguez *et al.*, 2013). Application of this approach to coherent X-ray diffraction imaging of single particles has been amply demonstrated (see, for example, Spence, 2008).

The problem is usually formulated as that of finding a point in the intersection of two constraint sets: the set of all functions that have the measured Fourier amplitude and the set of all functions that are contained within a given finite-extent region termed the 'support'. The function in our case is the electron density of the biological macromolecule under study. Such a point then satisfies the data and the known real-space constraints and is therefore a solution to the phase-retrieval problem.

Iterative projection algorithms (IPAs) are designed to tackle these constraint satisfaction problems. IPAs seek out the intersection between two constraint sets by iteratively searching through the multi-dimensional space that the problem resides in, using operators called projections.

It is convenient to formulate IPAs as operations on vectors in an n-dimensional metric space. A vector $\mathbf{f} = (f_1, \cdots, f_n)$ in this space represents the sampled electron density, where each of its n components f_i corresponds to the electron density at grid point i. A projection P_A is defined as an operation that takes a vector in this metric space to the closest vector in the constraint set A in this space. For example, the support constraint in phase retrieval requires the density in question to be zero outside the molecular support region S. The projection operator, P_S, that achieves this sets all values of the density outside the support to zero and leaves the values inside the support unchanged, *i.e.*

$$P_S\mathbf{f} = \begin{cases} f_i & \text{if } i \in S \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

Similarly, in reciprocal space, the projection operator denoted P_M sets the amplitude of the complex molecular transform equal to the measured amplitude (the set of all densities with Fourier amplitudes equal to the measured values is denoted M) whilst leaving the phase unchanged. The reciprocal-space amplitudes that are not measured are left at their current values during the application of the projection so they are free to change as the algorithm proceeds and are said to 'float'. The set of complex numbers that have the same amplitude defines a circle in the complex plane so that, geometrically, the projection involves moving the point radially to the closest point on that circle. The Fourier transform and inverse Fourier transform operations required to move the point back and forth between real space and reciprocal space are incorporated into the projection operator P_M which is given by

$$P_M\mathbf{f} = \mathcal{F}^{-1}\begin{cases}(F_i/|F_i|)(I_i)^{1/2} & \text{if } i \in M\\ F_i & \text{otherwise,}\end{cases} \quad (9)$$

where $\mathcal{F}$ denotes the Fourier transform such that $\mathcal{F}\{\mathbf{f}\} = (F_1, \cdots, F_n)$ is the molecular transform on the grid points in reciprocal space, I_i is the value of the intensity data at grid point i and M denotes the set of available intensity data. The amplitudes $|F_i|$ not in M are left to float.

An IPA generates a sequence of points $\mathbf{f}_j$, starting at a random position in the metric space, that are used to locate the solution. Note that the points $\mathbf{f}_j$ are not themselves estimates of the electron density, but are used to find the electron density, and are referred to as the 'iterate'. At the jth iteration of an IPA, the current iterate $\mathbf{f}_j$ is updated to form the next iterate $\mathbf{f}_{j+1}$ using an update rule that is a combination of the projections P_S and P_M acting on $\mathbf{f}_j$. Different IPAs are distinguished by different update rules. An effective IPA is the difference map (DM) algorithm (Elser, 2003*a*), which has the update rule

$$\mathbf{f}_{j+1} = \mathbf{f}_j + \beta\big(P_S L_M \mathbf{f}_j - P_M L_S \mathbf{f}_j\big), \quad (10)$$

where β is a parameter, and the operators L_S and L_M are called relaxed projections, given by

$$\begin{aligned} L_S\mathbf{f}_j &= (1+\gamma_M)P_S\mathbf{f}_j - \gamma_M\mathbf{f}_j\\ L_M\mathbf{f}_j &= \big(1+\gamma_S\big)P_M\mathbf{f}_j - \gamma_S\mathbf{f}_j, \end{aligned} \quad (11)$$

where γ_S and γ_M (usually fixed) are called relaxation parameters. Good convergence is generally obtained by setting the relaxation parameters to $\gamma_S = -1/\beta$ and $\gamma_M = 1/\beta$ (Elser, 2003*a*) and the algorithm then has the single parameter β. A useful property of the DM algorithm is that once it converges, *i.e.* $\mathbf{f}_{j+1} = \mathbf{f}_j$, a solution that satisfies both constraints, $\hat{\mathbf{f}}$, can be obtained immediately as

$$\hat{\mathbf{f}} = P_S L_M \mathbf{f}_j = P_M L_S \mathbf{f}_j. \quad (12)$$

The DM algorithm has good global convergence properties and is the algorithm we have chosen to apply to the nanocrystallography phase-retrieval problem.

4. Noise amplification and its amelioration

The primary difficulty with direct phasing in nanocrystallography as proposed above is that the estimates of the molecular transform obtained between the Bragg reflections are highly noisy. This can be seen as follows. If we additively decompose the noisy measured average intensity into its noiseless component, $\langle I_n(\mathbf{u})\rangle_n$, and the (photon) noise contribution, noise($\mathbf{u}$), and denote by $\big|F(\mathbf{u})\big|_{\rm P}$ the estimated molecular transform magnitude that is used for phasing, then equation (3) is replaced by

$$\big\langle I_n(\mathbf{u})\big\rangle_n + \text{noise}(\mathbf{u}) = \big|F(\mathbf{u})\big|_{\rm P}^2 Q^2(\mathbf{u}). \quad (13)$$

The estimated molecular transform amplitude obtained from the data is then given by

$$\big|F(\mathbf{u})\big|_{\rm P}^2 = \big|F(\mathbf{u})\big|^2 + \frac{\text{noise}(\mathbf{u})}{Q^2(\mathbf{u})} \quad (14)$$

and the noise is therefore amplified in regions where $Q^2(\mathbf{u})$ is small, *i.e.* between the Bragg reflections. The statistics of this noise amplification are described by Chen *et al.* (2014). A key observation from that analysis is that the SNR at a sample position $\mathbf{u}_i$ in reciprocal space for the phasing intensity is proportional to the value of the averaged shape transform at that position, *i.e.*

$$\text{SNR}_{\text{P}i} \propto Q(\mathbf{u}_i), \quad (15)$$

where $\text{SNR}_{\text{P}i}$ is the SNR for the phasing intensity at position $\mathbf{u}_i$, calculated as the mean of the diffracted signal divided by the standard deviation of the noise. Note that the overall SNR of the whole derived data set, denoted $\text{SNR}_{\rm P}$, calculated by taking the mean of the signal mean over all reciprocal-space data positions divided by the mean of the standard deviation of the noise, again over all available reciprocal-space positions, is smaller than the overall SNR of the measured data set, denoted $\text{SNR}_{\rm M}$, as a result of the division by $Q^2(\mathbf{u})$. This deterioration of the SNR of the phasing amplitude relative to the measured amplitude can be quantified by the ratio

$$\frac{\text{SNR}_{\rm P}}{\text{SNR}_{\rm M}} = \frac{1}{p}\left[\sum_{i=1}^{p} Q^2(\mathbf{u}_i)\sum_{i=1}^{p}\frac{1}{Q^2(\mathbf{u}_i)}\right]^{-1/2} \quad (16)$$

and worsens (gets smaller) as the mean crystallite size increases (Chen *et al.*, 2014).

Since $Q(\mathbf{u}_i)$ has a wide dynamic range, so does $\text{SNR}_{\text{P}i}$, which needs to be considered in the phase-retrieval process. Furthermore, the averaged shape transform $Q^2(\mathbf{u})$, which is estimated from the data, gives an estimate of $\text{SNR}_{\text{P}i}$ for each datum i. Our objective, therefore, is to use this information to ameliorate, as much as possible, the deleterious effects of the variable SNR on the reconstructed electron density.

Since the SNR is spatially variable in reciprocal space, our strategy is to sample the data in such a way as to maximize the SNR of the data needed for phase retrieval. It is necessary to consider both the number of data used (to ensure that the problem remains well determined) and their positions in reciprocal space (to maximize the phasing SNR). For computational convenience, the diffraction amplitude data are sampled onto a grid in reciprocal space that is finer than the reciprocal lattice. If this grid oversamples reciprocal space by a factor s in each direction relative to the reciprocal lattice, then the oversampling factor O of the three-dimensional data set is defined as

$$O = s^3. \quad (17)$$

Since we are using an estimate of the continuous molecular transform to estimate the electron density, we effectively have a phase problem for a single particle (the contents of one unit cell). Under these circumstances, and under the assumption that the molecular support region is approximately convex and centrosymmetric, the minimum number of amplitude data required to uniquely define the electron density is twice the number of Bragg samples at the particular resolution of the

data (Miao *et al.*, 1998; Elser & Millane, 2008), *i.e.* uniqueness requires that $O > 2$. Our approach then is to retain a subset of the oversampled amplitude data with the largest SNR and use these for phase retrieval. The amplitudes at the remaining data points are determined by the real-space constraints, *i.e.* they are allowed to float. If, as a result of removing data with a low SNR, a proportion $0 \leq P \leq 1$ of the data are retained, then the oversampling factor of the data used for phasing, denoted $O_{\rm P}$, is

$$O_{\rm P} = PO = Ps^3. \tag{18}$$

Uniqueness of the solution then requires that $O_{\rm P} > 2$, and a margin on this inequality is desirable in practice.

Increasing the oversampling s increases the size of the computational grid by a factor s^3, so it is desirable to keep s as small as possible. Since $Q^2(\mathbf{u})$ provides an estimate of the SNR of each datum, our objective is to retain data points where $Q^2(\mathbf{u})$ is large. For $s = 2$, $O = 8$, but the additional data (*i.e.* in addition to the Bragg samples) are midway between the Bragg reflections where $Q^2(\mathbf{u})$ is smallest. This choice of s is therefore unsuitable. For $s = 3$, $O = 27$, the points where $Q^2(\mathbf{u})$ is smallest are avoided, and O is large enough that a significant proportion of the low-SNR data can be removed while still satisfying $O_{\rm P} > 2$. For $s = 4$, $O = 64$, and some of the new samples fall where $Q^2(\mathbf{u})$ is smallest. Given the additional computational cost for $s = 4$, this value does not appear to offer any advantages over $s = 3$. Therefore, $s = 3$ appears to be a suitable value. Note that although increasing s increases the number of data, the data points become closer together and therefore the data become more correlated and less information is added.

Since the SNR is proportional to $Q^2(\mathbf{u})$, a sensible approach is to set a threshold, ξ, on $Q^2(\mathbf{u})$ and to use the data for which $Q^2(\mathbf{u}) > \xi$. The threshold determines the proportion P of the data that are retained and must be chosen such that $O_{\rm P}$ is sufficiently large. The effect of the threshold is illustrated in one dimension in Fig. 1. The samples that are not used as data

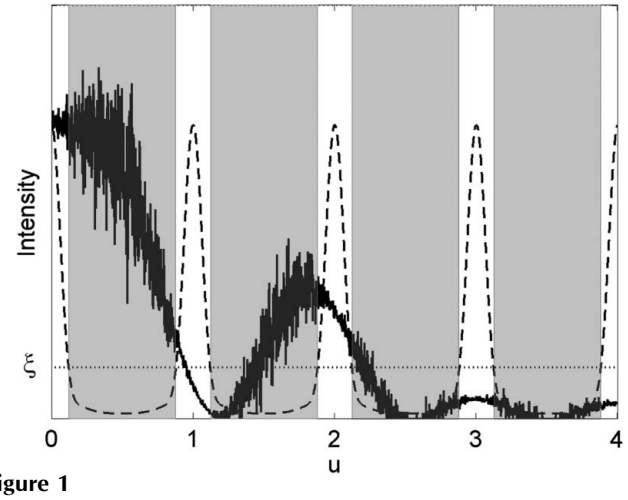

Figure 1
A threshold (dotted line) on the averaged shape transform (dashed line) determines the regions (gray) where the intensity data are not used. The solid line shows the noisy estimate of the molecular transform.

Table 1
Oversampling factor $O_{\rm P}$ and the proportion of data retained P for the four sampling schemes for $s = 3$ ($O = 27$).

Sampling scheme	$O_{\rm P}$	P
A	27	1.00
B	19	0.70
C	7	0.26
D	1	0.04

(gray region in Fig. 1) are treated as missing data and their values are made to float during the phase-retrieval process. The threshold value is normalized such that it is unity at the Bragg peaks.

For example, for the case of an orthorhombic unit cell with a crystallite size distribution with the same marginal density in the three directions, as a result of the symmetry of $Q^2(\mathbf{u})$ and of the reciprocal lattice, for $s = 3$ there are four possible sampling schemes corresponding to four values of ξ. These sampling schemes are labelled A, B, C, D in order of increasing ξ and their oversampling factors are listed in Table 1. Scheme A corresponds to maximal oversampling (no data removed) and scheme D corresponds to no oversampling (Bragg samples only). Scheme C includes the Bragg samples and the six samples closest to the Bragg samples on lines parallel to the reciprocal-space axes. Scheme B excludes the 'body diagonal' additional samples from the full set of oversampled data of scheme A. We note that in general $Q^2(\mathbf{u})$ will tend to decrease monotonically with distance from the Bragg samples so that in practice it will likely be sufficient, and more convenient, to select sampling schemes based on this distance rather than a threshold on the measured $Q^2(\mathbf{u})$.

5. Simulations

The selective sampling approach described in §4 was tested by implementing different sampling schemes and retrieving the phases of simulated diffraction data using the DM algorithm. The simulations were conducted in three dimensions. The molecule used for the simulations was the membrane protein aquaporin 1 (AQP1) (Ren *et al.*, 2000), for which a $32 \times 32 \times 45$ Å volume of the electron density, sampled on a 1 Å grid, was used. Reciprocal space was oversampled by a factor of three in each direction. This was done by zero-padding the real-space volume out to $96 \times 96 \times 135$ grid points prior to calculating the Fourier transform. The true diffracted intensities were calculated and corrupted with Poisson noise. The noise level on the simulated data was manipulated as follows. Letting Po(λ) denote the Poisson distribution with parameter λ and I the original noiseless value of the intensity, the corrupted intensity is calculated as

$$I_{\rm noisy} = \frac{1}{k} I', \tag{19}$$

where the scaling factor k is a parameter that controls the noise level and I' is drawn from the distribution Po(kI). Thus the larger the value of k, the smaller the Poisson noise and the

larger the SNR. The SNR of a Poisson random process is the square root of the mean, so that in this case the SNR of the sample intensity is $(kI)^{1/2}$. The noiseless case can be thought of as having a scaling factor k of infinity.

A Gaussian crystal size distribution is used with the same mean number of unit cells in each direction, denoted μ_N, and a standard deviation $\sigma_N = \mu_N/3$. This ensures that the probability of one side of a crystal being less than one unit cell is negligible and the distribution is truncated at $N = 1$. The number of unit cells in each of the three directions are assumed to be independent. For each value of μ_N considered, the averaged shape transform $Q^2(\mathbf{u})$ was calculated using equation (5). The noisy amplitudes computed as described above were divided by this averaged shape transform to calculate the amplitudes to be used for phasing. Intensity samples within a sphere of radius of five grid points centered at the origin of reciprocal space were discarded to simulate the effect of the X-ray beam stop. The intensity samples outside a sphere of radius 0.5 Å^{-1} in reciprocal space were also not used, resulting in a resolution of the reconstructed electron density of 2 Å. The high resolution limit reduces the number of data and hence the oversampling factor, by a factor of about two.

The DM algorithm as described in §3 was used to reconstruct the electron density for different average crystal sizes and noise levels. The algorithm parameter β was set to 0.7. Constraints imposed in real space are the support constraint being the same size as the density within the zero-padded array, and reality of the electron density. The constraint in reciprocal space is that the amplitudes are equal to the data values except for those samples outside the resolution sphere, around the origin of reciprocal space within the obscured zone of the beam stop, and those which are excluded by the particular sampling scheme used. The amplitudes in the excluded regions are allowed to float.

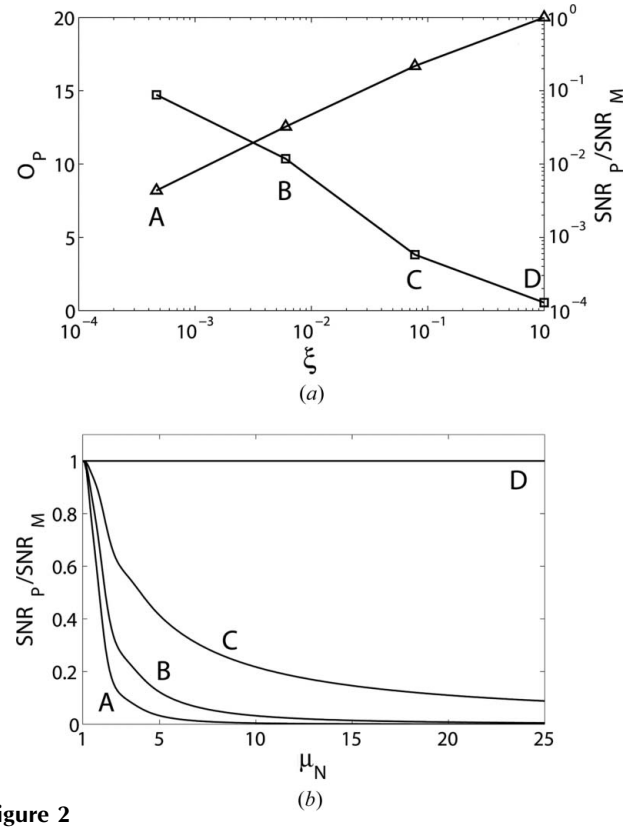

Figure 2
(*a*) The oversampling factor O_P (squares) and the SNR ratio (triangles), for sampling schemes A, B, C and D, with $\mu_N = 10$. (*b*) The SNR ratio for the four sampling schemes as a function of mean crystallite size.

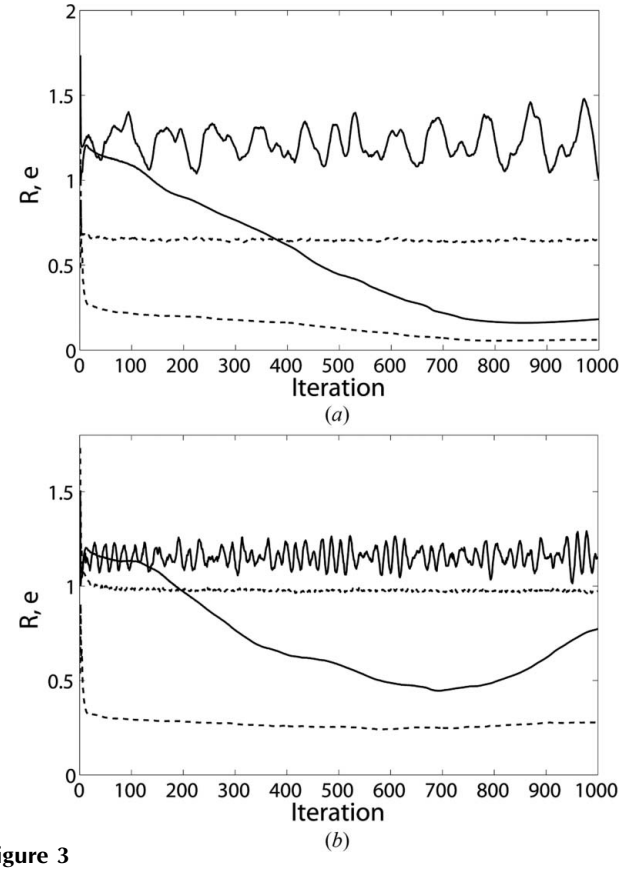

Figure 3
R factor (dotted line) and r.m.s. error (solid line) in the electron density *versus* iteration using sampling schemes A (upper curves at the 1000th iteration) and C (lower curves at the 1000th iteration), for (*a*) $\mathrm{SNR_M} = 100$ and (*b*) $\mathrm{SNR_M} = 20$.

Progress of the algorithm is monitored by calculating the crystallographic *R* factor

$$R = \frac{\sum_{\mathbf{u}} \left| |\hat{F}(\mathbf{u})| - |F(\mathbf{u})|_{\mathrm{P}} \right|}{\sum_{\mathbf{u}} |F(\mathbf{u})|_{\mathrm{P}}} \tag{20}$$

as a function of iteration, where $|\hat{F}(\mathbf{u})|$ is the Fourier magnitude of the estimated solution $\hat{f}(\mathbf{x})$ obtained using equation (12) where $\mathbf{x}$ is the position in real space and $|F(\mathbf{u})|_{\mathrm{P}}$ is the Fourier magnitude data. The quality of the reconstruction is measured by calculating the root-mean-squared (r.m.s.) error in the reconstructed electron density

$$e = \frac{\|\hat{\mathbf{f}} - \mathbf{f}\|}{\|\mathbf{f}\|} = \left\{ \frac{\sum_{\mathbf{x}} \left[\hat{f}(\mathbf{x}) - f(\mathbf{x}) \right]^2}{\sum_{\mathbf{x}} f^2(\mathbf{x})} \right\}^{1/2} \tag{21}$$

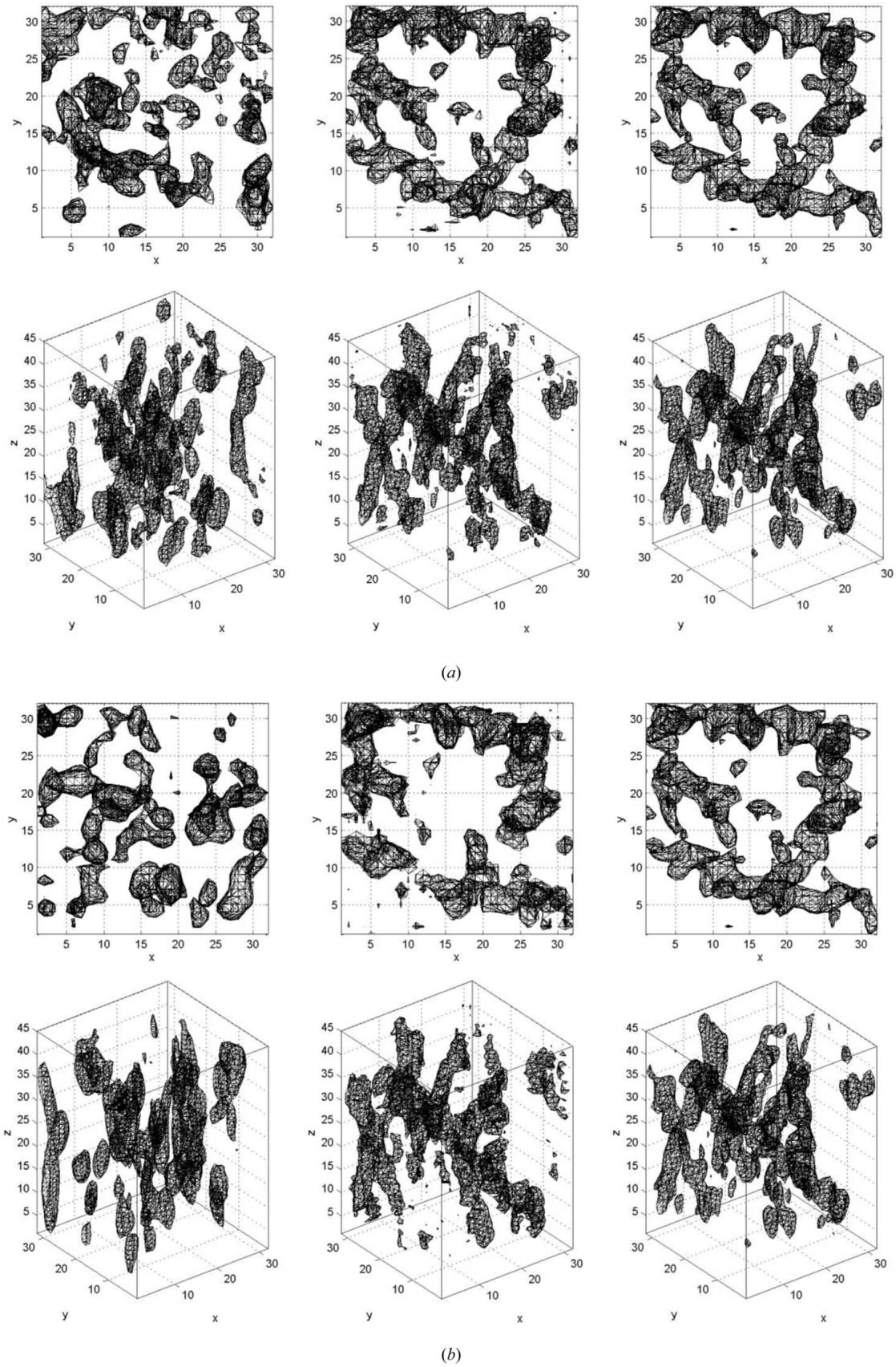

Figure 4
Reconstructed electron densities for (*a*) $\mathrm{SNR_M} = 100$ and (*b*) $\mathrm{SNR_M} = 20$. The top rows show the projected views along the z axis and the bottom rows the three-dimensional volume. The left and center columns are reconstructions using sampling schemes A and C, respectively, and the right column is the true electron density.

as a function of iteration, where $\|\cdot\|$ is the Euclidean norm, and $\mathbf{f}$ and $f(\mathbf{x})$ denote the true electron density.

6. Results

Simulations were conducted for the two values $\mathrm{SNR_M} = 100$ and 20. An oversampling factor $s = 3$ in each direction and a mean crystal size $\mu_N = 10$ were used, and reconstructions calculated using the four sampling schemes listed in Table 1. The DM algorithm was run for 1000 iterations, starting with a random electron density. The final reconstructed electron density was chosen as that with the minimum value of R.

The oversampling factor O_P and the expected SNR ratio $\mathrm{SNR_P}/\mathrm{SNR_M}$ calculated using equation (16) for the four sampling schemes are shown in Fig. 2(*a*). Note that the values of O_P in Fig. 2(*a*) are smaller than those in Table 1 as a result of removal of the high-resolution data as described in §5. Inspection of Fig. 2(*a*) shows that the best SNR with sufficient oversampling factor ($O > 2$) is obtained with sampling scheme C. The overall SNR ratio for the four sampling schemes is shown *versus* mean crystal size in Fig. 2(*b*). This shows the deterioration in $\mathrm{SNR_P}$ relative to $\mathrm{SNR_M}$ with increasing crystal size as noted previously (Chen *et al.*, 2014) and also the improvement in $\mathrm{SNR_P}$ for sampling schemes B and C over using all samples (scheme A). Note that, for a fixed incident X-ray pulse flux, $\mathrm{SNR_M}$ will increase with increasing crystallite size.

For the simulated reconstructions, the R factor and r.m.s. error *versus* iteration of the algorithm for sampling schemes A and C are shown in Fig. 3. For both SNRs, the algorithm converges to a low r.m.s. error for the selective sampling scheme C but not when all samples are used (sampling scheme A). This shows the advantage of selecting the data with the best SNR. The resulting reconstructed electron densities are shown in the left and center columns of Fig. 4, where they are compared to the true electron density (right column). It is clear that using the selective sampling leads to an interpretable density whereas using all the data does not.

The R factor and electron-density r.m.s. error for the final reconstructions using the four sampling schemes A–D are shown in Fig. 5, where it is seen that sampling scheme C does

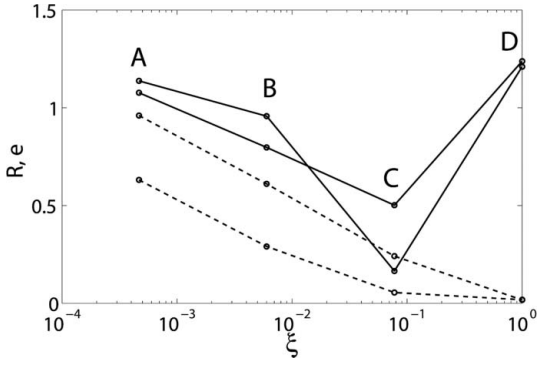

Figure 5
R factor (dotted line) and electron density r.m.s. error (solid line) of the final reconstructions for sampling schemes A–D for $\mathrm{SNR_M} = 100$ (lower curves at C) and $\mathrm{SNR_M} = 20$ (upper curves at C).

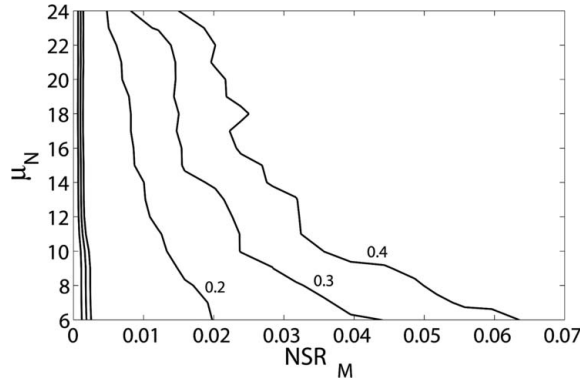

Figure 6
Contour plots of constant r.m.s. error in the reconstructed electron density for sampling scheme A (left three curves) and sampling scheme C (right three curves), *versus* mean crystal size and noise-to-signal ratio for the measured data $\mathrm{NSR_M}$. The r.m.s. error is contoured at 0.2, 0.3 and 0.4 for both sampling schemes.

indeed give the best reconstruction. Referring to Figs. 2(*a*) and 5, the reconstructions improve as the SNR ratio increases and data are removed until the point at which there are not enough data, at sampling scheme D, where $O < 2$. Note that for sampling scheme D a small R factor is obtained since the problem is under-constrained and the algorithm finds one of a multitude of incorrect solutions that satisfy the data.

For a fixed $\mathrm{SNR_M}$, $\mathrm{SNR_P}$ deteriorates as the mean crystal size of the nanocrystals increases. Simulations were conducted for a range of noise levels in the data and crystallite sizes. The results are shown in Fig. 6, which shows the error in the final reconstructions *versus* the noise level in the data (noise-to-signal ratio $\mathrm{NSR_M} = 1/\mathrm{SNR_M}$) and the mean crystal size for sampling schemes A and C. Inspection of Fig. 6 shows that for a particular crystal size, significantly larger noise levels can be tolerated by using the sample-selection scheme.

7. Discussion

Nanocrystallography using X-ray free-electron lasers offers the possibility of direct phasing of the diffraction data using measurement of the diffracted intensity between the Bragg reflections to estimate the molecular transform. This estimate is noisy however at positions between the Bragg reflections. To address this problem, a selective sampling strategy that retains only the measured intensity samples that have a highest signal-to-noise ratio is employed. The averaged shape transform that is estimated from the diffraction data can be used to determine this sampling scheme, although in practice a scheme based on the distance of samples from the Bragg reflections is probably sufficient. Oversampling the reciprocal lattice by a factor of three in each direction allows removal of low-SNR data while retaining sufficient data for a unique solution and minimizing the computational load. Simulations show that using this selective sampling with the difference map algorithm allows reconstruction at lower SNR than if all the data are used. The results show the trade-off between noise level and crystallite size that can be tolerated for direct phasing in nanocrystallography.

This work was supported by a James Cook Research Fellowship to RPM, an NSF grant (MCB-1021557) to JCHS, NSF STC award 1231306, and a UC Doctoral Scholarship to JPJC. The authors would like to thank Phil Bones for helpful discussions and Alok Mitra for providing the AQP1 electron-density map.

References

Barty, A. *et al.* (2011). *Nat. Photon.* **6**, 35–40.
Bates, R. & McDonnell, M. (1989). *Image Restoration and Reconstruction.* New York: Oxford University Press.
Chen, J. P. J. & Millane, R. P. (2013). *J. Opt. Soc. Am. A*, **12**, 2627–2634.
Chen, J. P. J., Spence, J. C. H. & Millane, R. P. (2014). *Acta Cryst.* A**70**, 143–153.
Dilanian, R. A., Streltsov, V. A., Quiney, H. M. & Nugent, K. A. (2013). *Acta Cryst.* A**69**, 108–118.
Elser, V. (2003*a*). *J. Opt. Soc. Am. A*, **20**, 40–55.
Elser, V. (2003*b*). *Acta Cryst.* A**59**, 201–209.
Elser, V. & Millane, R. P. (2008). *Acta Cryst.* A**64**, 273–279.
Fienup, J. R. (1982). *Appl. Opt.* **21**, 2758–2769.
Kirian, R. A., Wang, X., 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.
Liu, H., Zatsepin, N. A. & Spence, J. C. H. (2014). *IUCrJ*, **1**, 19–27.
Martin, A. V. *et al.* (2012). *Opt. Express*, **20**, 16650–16661.
Miao, J. *et al.* (1999). *Nature (London)*, **400**, 342–344.
Miao, J. & Sayre, D. (2000). *Acta Cryst.* A**56**, 596–605.
Miao, J., Sayre, D. & Chapman, H. N. (1998). *J. Opt. Soc. Am. A*, **15**, 1662–1669.
Millane, R. P. (1990). *J. Opt. Soc. Am. A*, **7**, 394–411.
Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. (2000). *Nature (London)*, **406**, 752–757.
Redecke, L. *et al.* (2013). *Science*, **339**, 227–230.
Ren, G., Cheng, A., Reddy, V., Melnyk, P. & Mitra, A. K. (2000). *J. Mol. Biol.* **301**, 369–387.
Rodriguez, J. A., Xu, R., Chen, C.-C., Zou, Y. & Miao, J. (2013). *J. Appl. Cryst.* **46**, 312–318.
Sayre, D. (1952). *Acta Cryst.* **5**, 843.
Son, S. K., Chapman, H. N. & Santra, R. (2011). *Phys. Rev. Lett.* **107**, 218102.
Spence, J. C. H. (2008). *Diffractive (Lensless) Imaging*, ch. 19. New York: Springer.
Spence, J. C., Kirian, R. A., Wang, X., Weierstall, U., Schmidt, K. E., White, T., Barty, A., Chapman, H. N., Marchesini, S. & Holton, J. (2011). *Opt. Express*, **19**, 2866–2873.
Vartanyants, I. A. & Robinson, I. K. (2001). *J. Phys. Condens. Matter*, **13**, 10593–10611.