Phase retrieval for diffraction data from very small crystals

Joe Chen and Rick Millane
Computational Imaging Group, Department of Electrical and Computer Engineering,
University of Canterbury, Christchurch, New Zealand
Email: rick.millane@canterbury.ac.nz

***Abstract*—Nanocrystallography is a form of coherent X-ray diffraction imaging that utilises a stream of small crystals of the biological assembly under study. These small crystallites are termed nanocrystals in contrast to the large crystal formations used in conventional X-ray crystallography. Diffraction snapshots of individual nanocrystals can be obtained using femtosecond pulses from an X-ray free-electron laser (XFEL). A key advantage of nanocrystallography is that many membrane proteins for example do not easily form large crystals, but do form nanocrystals and thus crystallography techniques can be applied to image these previously unapproachable structures. In this paper we present results from a study of the performance of phase retrieval in the context of nanocrystals. We use a variable signal-to-noise ratio (SNR) determined by a shape transform function averaged over a distribution of crystallite sizes and investigate convergence of the difference map algorithm as a function of the overall SNR. The results give a picture of the noise levels and crystallite sizes that can be tolerated in nanocrystallography.**

I. Introduction

X-ray crystallography is a technique for imaging single biological molecules using X-rays diffracted from a regular formation of the molecule under study. This well-ordered molecular arrangement is referred to as a crystal. Since the diffracted X-ray amplitude is the Fourier transform of the electron density in the crystal (from which the positions of the atoms inside the molecule can be inferred), in principle this electron density can be recovered by the inverse Fourier transformation. There are three difficulties however. First, only the amplitude, but not the phase, of the diffracted X-rays can be measured - this is an example of what is often called a "phase problem" where the Fourier phase information is lost and need to be retrieved. Second, since the crystal is periodic by nature, its Fourier transform, i.e. the diffracted X-ray amplitudes, are discrete in Fourier space and we observe the so-called Bragg reflections or Bragg peaks. These peaks undersample the amplitude of the diffraction pattern and so the phase problem is underdetermined. The third difficulty of X-ray crystallography is simply that for some molecular assemblies the sample of interest cannot be easily crystallised.

The first obstacle has been readily addressed through various techniques developed over the years to recover the phases from a measured diffraction amplitude. Empirical methods include the so-called molecular replacement and multiple isomorphous replacement processes where the molecule specimens themselves are physically modified [1]. Algorithmic methods include iterative projection algorithms (IPAs) that are used to numerically recover the phases. IPAs are what we are interested in here and will be elaborated on in section III.

The second and third difficulty of coherent X-ray diffraction imaging: the undersampling of the diffraction pattern and the crystallisation problem, have spurred on much of the effort to extend traditional crystallography techniques to crystals with a smaller number of unit cells, termed "nanocrystals". In contrast to larger crystals with many unit cells, small crystals offer the benefit of avoiding the need to crystallise the molecular sample in question, and gets around the problem of being able to measure only discrete Fourier amplitude samples by providing a glimpse of the continuous diffraction pattern from a single molecule.

It is not until very recently that the imaging of these nanocrystals can be practically achieved. This is because without the repetition of large numbers of unit cells from conventional crystals, the diffracted signal levels drop significantly as the intensity of the diffraction pattern is proportional to the total number of unit cells in the target crystal. Thus assuming fixed unit cell size, the smaller the crystal, the smaller the number of unit cells and therefore the weaker the diffracted signal which in turn means that noise and other parasitic effects can easily overwhelm the detection process.

The simplest way to overcome this problem is to increase the power of the incident X-ray beam. This will result in a higher photon count on the detector and give a better signal-to-noise ratio (SNR). However in doing so, the biological specimen can sustain severe radiation damage, resulting in a change in their intrinsic structure; or worse, their destruction.

The newly developed synchrotron technology provides a solution to this conundrum. Intense, ultrafast X-ray pulses can be created in accelerator facilities such as the Linac Coherent Light Source (LCLS) at the Stanford accelerator center in the US. These so-called X-ray free-electron lasers (XFEL) generate pulses of X-rays lasting only around 100fs but have an average of 10^{12} photons per pulse [2]. With these intense femtosecond X-ray pulses, appropriate signal levels can be obtained for nanocrystals and radiation damage on the biological specimen can be vastly reduced or even outrunned all together [3] because of the fleeting nature of the illumination.

Here we explore the consequences of coherent X-ray imaging under the framework of nanocrystals; in particular the effect on the process of retrieving the Fourier phases for

diffraction data generated from very small crystals of the order of less than a hundred unit cells. Implication of the need to use a collection of nanocrystals on the imaging process is first mathematically outlined and the projection algorithm used to retrieve the phases for reconstructing the image introduced. Lastly, results from computer simulations of the phase retrieval process of an actual protein are presented.

II. Diffraction by a collection of nanocrystals

As mentioned earlier, by impacting an object with a beam of incident radiation, the diffraction pattern registered on a detector is the Fourier transform of that object's electron density. For crystallised objects, the periodic repetition of the molecule of interest inside the crystal lattice means that the diffracted wavefronts will no longer represent the Fourier transform of the molecule itself but will be the molecule's transform modulated by a function specific to the crystal called its shape function.

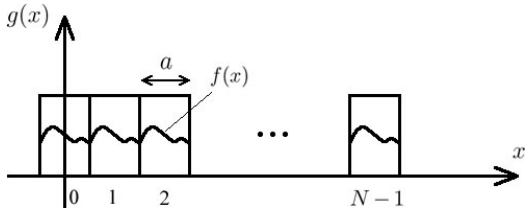


Fig. 1. Schematic view of a 1-dimensional nanocrystal.

Consider the 1-dimensional case where the molecule's density is denoted by $f(x)$ and is repeated within unit cells of width a, $N-1$ times along the x direction as shown in figure 1. The result is a 1-dimensional crystal with a density $g(x)$ composed of repeated copies of $f(x)$ which can be expressed as a convolution with a train of delta functions,

$$g(x) = f(x) \otimes \sum_{h=0}^{N-1} \delta(x - ha).$$

By hitting this crystal with X-rays and measuring the resulting diffracted radiation, we obtain the Fourier transform of $g(x)$; namely, $G(u)$. After jumping through some algebraic hurdles by evaluating the Fourier integral and using the formula for the geometric series, we can arrive at an expression for the transform of the crystal,

$$G(u) = F(u)\frac{\sin(\pi auN)}{\sin(\pi au)} \cdot e^{-j\pi au(N-1)} \tag{1}$$

where $F(u)$ is what we are interested in - the transform of the molecule in question; and the function that $F(u)$ gets multiplied with is precisely the shape function of the crystal, $S_N(u)$, giving us

$$G(u) = F(u)S_N(u). \tag{2}$$

We are only concerned with the magnitude of equation (1) as that is what we measure in practice, so we have

$$|S_N(u)| = \frac{\sin(\pi auN)}{\sin(\pi au)}. \tag{3}$$

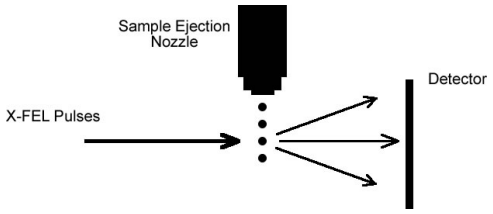


Fig. 2. Experimental set-up of diffraction pattern measurements from a stream of nanocyrstals.

In practice, nanocrystals of the biological sample of interest are delivered to the high intensity femtosecond X-ray pulses through an ejector nozzle as shown in figure 2. This nozzle is vibrated with an ultrasonic piezoelectric transducer which synchronises the ejecting crystals to a predetermined frequency; allowing each pulse of the femtosecond X-ray to strike one nanocrystal only [4]. By taking "snapshots" of each ejecting crystal, a collection of diffraction patterns are thus obtained on the same crystal and can be averaged to generate the Fourier magnitude data. The intensity of a single diffraction pattern is the squared magnitude of the transform of the crystal, $|G(u)|^2$. If we denote $P(N)$ as the probability distribution function for the sizes of the ejected nanocrystals from the molecular output nozzle, then the averaged diffraction intensity over a large collection of individual diffraction patterns is

$$I(u) = \sum_{\text{all N}} P(N)\,|G(u)|^2\,.$$

Upon substituting in the expression for the shape function and the molecular transform from equation (2) and (3) in place of $G(u)$ and rearranging, we obtain

$$I(u) = |F(u)|^2 \sum_{\text{all N}} P(N)\left(\frac{\sin(\pi auN)}{\sin(\pi au)}\right)^2.$$

The summation term over all crystal sizes represents the modulation term altering the intensities of the diffraction pattern. If we let $Q_N(u)$ to be the square root this modulation term, we have the averaged shape function of a collection of nanocrystals,

$$Q_N(u) = \sqrt{\sum_{\text{all N}} P(N)\left(\frac{\sin(\pi auN)}{\sin(\pi au)}\right)^2}. \tag{4}$$

Thus to obtain the Fourier magnitudes needed for phase retrieval, we need to divide the square root of the measured intensity by $Q_N(u)$. The measured intensity in practice will always contain noise and so the measured magnitude of the molecule's transform is

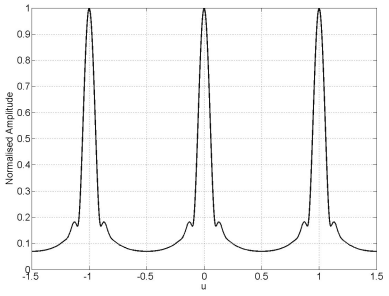
$$|F(u)|_{\text{meas}} = |F(u)| + \frac{\text{noise}}{Q_N(u)}, \tag{5}$$

assuming additive noise. The inverse of $Q_N(u)$ thus becomes the function that determines the amplification of the measurement noise.

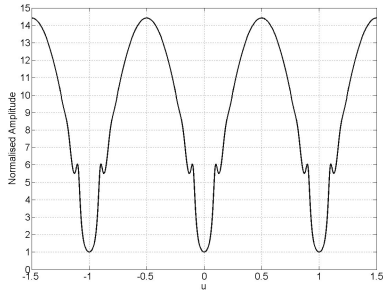
Figure 3(a) shows the normalised amplitudes of the averaged shape function for a collection of nanocrystals with a

Gaussian size distribution. The mean crystallite size is 10 unit cells and the standard deviation is 2 unit cells. It can be seen that the peaks of $Q_N(u)$ centre around integer multiples of the inverse unit cell width, $1/a$, where $a = 1$ in this case. These points are known as the reciprocal lattice points and this is what gives rise to the Bragg peaks seen in measured diffraction patterns. As the nanocrystals' sizes increase, these peaks become concentrated at the lattice points and the result is a diffraction pattern of a discrete nature. The data between the Bragg peaks are lost for large crystals because of the disproportionate amplification of noise as evident from the large amplitudes seen in the plot of the inverse averaged shape function in figure 3(b).

Thus the sizes of the nanocrystals and their distribution determines the level of noise corruption in the measurement data, which in turn poses a significant influence on the process of retrieving the Fourier phases in order to reconstruct the image of the molecule.



(a)



(b)

Fig. 3. The averaged shape function and its inverse for a normally distributed collection of nanocrystals with a mean crystallite size of 10 and a standard deviation of 2. (a) and (b) shows the normalised amplitude of $Q_N(u)$ and $1/Q_N(u)$ respectively.

III. Phase retrieval

To retrieve the Fourier phases when only the magnitudes are known is by itself not possible to do in principle as these two quantities are independent of one another. However we can utilise further information specific to the problem at hand. For example the knowledge that all objects in practice are non-negative and of finite extent helps a great deal in constraining the problem to allow us to arrive at a unique solution of the phases. The formulation of the problem then becomes that of trying to find the intersection between two constraint sets: the set of all possible objects that have the measured Fourier modulus and the set of all possible objects that are contained within the given finite-extent region termed the support. The object here in our case is the electron density of the biological molecule in question. In this paper, the terms "objects", "images" and "densities" shall be used interchangeably and assumed to have the same meaning.

Iterative projection algorithms (IPAs) are specifically designed to tackle these constraint satisfaction problems. They 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 and depending on the shape of the constraints in this space (termed their convexity) an intersection can usually be found if one exists.

Gerchberg and Saxton proposed one of the earliest iterative algorithms to retrieve phases of complex objects in a paper in 1972 [5]. Fienup later refined and extended this technique to propose the so-called error reduction algorithm (ER) and the hybrid-input-output algorithm (HIO) in 1978 [6]. The former is the simplest conceptual form of IPA similar to what Gerchberg and Saxton had in mind except real objects are used and the constraint in the object domain is the finite-extent criterion; the latter is based on methods using nonlinear feedback control theory [7].

It is convenient to formulate IPAs as operations on vectors in an n-dimensional metric space. A vector $x = (x_1, x_2, \cdots, x_n)$ in this space represents an n-pixel image where each of its n components correspond to the value of a particular pixel. These n coordinates form an orthogonal basis for the n-dimensional space that they reside in; an argument which holds since all pixels of an image can be assumed to be independent.

A projection P_A is then defined as an operation that takes a point in this metric space to the closest point on some set A in this space. For example, the support constraint in phase retrieval requires the density of the object in question to be zero outside a support region. The projection operator, P_S, that achieves this is obtained by setting all values of the image outside the support to zero; or in terms of our metric space formalism, zeroing the value of those components in the n-dimensional vector outside the support region S.

$$P_S x = \begin{cases} x_i & \text{if } x_i \in S \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, in the Fourier domain, projection P_M sets the magnitude of a complex number to that of the measured magnitude whilst leaving the phases unchanged. The set of all complex numbers that have the same magnitude defines a circle on the complex plane, therefore the projection involves moving the target point radially to the closest point on the circle. In practice, the Fourier transform and the inverse Fourier transform required to move the point back-and-forth between the object domain and the frequency domain is incorporated into the P_M projection operator itself.

At the nth iteration of an IPA, the current iterate x_n is updated to form the next iterate x_{n+1}. The update rule of the algorithm is a combination of the projections P_S, P_M and the identity operator I, where different IPAs are distinguished by different update rules. The ER algorithm is when the two projections are simply applied sequentially to x_n, giving us

$$x_{n+1} = P_S P_M x_n. \tag{6}$$

This algorithm has the desirable property that given the constraint set is convex, the error always decreases monotonically with increasing numbers of iteration [6]; hence its given name. However the problem with the ER algorithm is that it is prone to become trapped inside false fixed points that do not satisfy both constraints. Fienup's HIO algorithm takes the form

$$x_{n+1} = \begin{cases} P_M x_n & \text{if } x \in S \\ (I - \beta P_M)\, x_n & \text{otherwise,} \end{cases}$$

and is less susceptible to being trapped in false fixed points.

One of the most effective IPAs to date was proposed by Elser at the turn of this century [8]. This so called difference map (DM) algorithm involves the use of three independent parameters γ_S, γ_M and β and takes the form

$$x_{n+1} = \big(I + \beta\,(P_S F_M - P_M F_S)\big) x_n \tag{7}$$

where

$$F_S = (1 + \gamma_M)\, P_S - \gamma_M I,$$

$$F_M = (1 + \gamma_S)\, P_M - \gamma_S I.$$

F_S and F_M are called "relaxed projections" as they do not move onto the closest point in the respective constraint sets but rather to a point slightly beyond the constraints. Elser suggested as possible values for optimum performance, the values $\gamma_S = -1/\beta$ and $\gamma_M = 1/\beta$. For the DM algorithm the iterate itself is not an estimate of the solution and an estimate of the solution x_{soln} can be obtained as

$$x_{\text{soln}} = P_S F_M x_n.$$

The DM algorithm has good global convergence properties and is thus the algorithm we chose to apply to our nanocrystallography phase retrieval problem at hand.

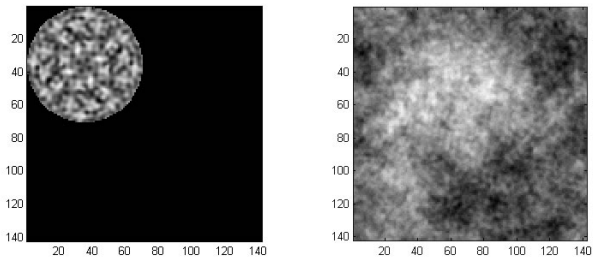
IV. Simulation Setup

A 2-dimensional image of the circular protein erythrocyte aquaporin 1 (AQP1) from [9] is used as the test object in our nanocrystallography phase retrieval simulations in MATLAB where the image is converted to 8-bit greyscale with pixel values between 0 and 255.

The Fourier transform of the 71 pixel by 71 pixel image is first obtained and is used as the true magnitudes that would be measured in an ideal noiseless world. To simulate the diffraction data measured in practice, the true magnitudes are corrupted by an additive Gaussian noise with zero mean and amplified by the inverse averaged shape function in 2-dimensions as required by equation (5).

As mentioned earlier, using diffraction data only at the Bragg peaks leads to an underdetermined situation where there are more unknowns than equations. Therefore we must oversample the transform to ensure sufficient information is obtained and this can be done by padding the image with zeros to obtain a new image that is twice the size as the original as shown in figure 4(a). The image size is doubled because the minimum oversampling factor needed for a unique reconstruction is two, as first shown by Sayre [10] in 1952 using Nyquist-Shannon sampling arguments. The consequence of this is that each pixel of the transformed, double-sized image is lying on integer multiples of half the Bragg peak spacings and thus the noise that gets added onto those pixels also gets amplified according to $1/Q_N(u)$ at those spacings.

The phase of the transform of this enlarged image is pretend not to be known and is replaced with the phases from the transform of a random image of the same size. This forms the starting iterate to be fed into our DM algorithm and is shown in figure 4(b).



(a) (b)

Fig. 4. Simulation set-up. (a) shows the zero-padded image of double the size of the original image in both dimensions and (b) shows the resulting starting iterate where the magnitude of the true image is mixed with random phase.

The mask for the support projection of the DM algorithm is made to be a circle that is slightly larger than the circular image of the protein itself. The support is loose because it was found that with a tight support, i.e. the mask exactly equal to the outline of the protein, the errors in the reconstruction diverges due to circular shifts of the Fourier transform being introduced in the reconstruction process where the pixels at the boundary of the image wraps around to the other side.

The nanocrystals in our simulation are assumed to be 2-dimensional and square. A normal distribution with a standard deviation one-fifth of the mean is used as the probability distribution for the sizes of the nanocrystals ejecting from the output nozzle. So for example, a set of crystals with a mean size of 10 unit cells in both directions has a standard deviation of 2 unit cells also in both directions. Having a standard deviation that varies with the mean is a reasonable assumption as the smaller the mean crystallite size, the smaller its deviation should be. The direct current portion of the Fourier magnitude is also discarded to simulate the effect of the X-ray beam-stop.

V. Results

Upon application of the DM algorithm, the result of simulating a collection of 2-dimensional nanocrystals with a mean crystallite size of 10 after 1000 iterations is shown in figure 5. Two metrics were used to gauge the quality of the reconstruction. The first is an error metric designed for the DM algorithm which is simply the difference between the next iterate and the current iterate, denoted by $\Delta = ||x_{n+1} - x_n||$ where the norm used in this case is the Euclidean norm. Small values of Δ signals the convergence of the iterate and the consequent arrival at a fixed point.

The second metric is the so-called R-factor and is the standard quantity used in crystallography to measure the errors of the reconstructed image in which the true solution is not known. It is defined as

$$\frac{\Sigma_u \big| \, |F(u)|_{\text{meas}} - |\hat{F}(u)| \big|}{\Sigma_u \, |F(u)|_{\text{meas}}}$$

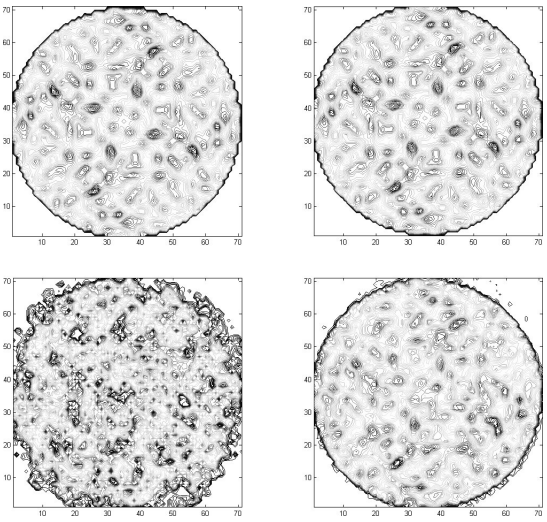
where $|F(u)|_{\text{meas}}$ is the measured amplitude and $|\hat{F}(u)|$ is the magnitude of the estimate of the solution. In general, an R-factor of less than 0.3 indicates an acceptable reconstruction has been obtained.

The top right-hand corner of figure 5(a) shows the reconstruction for the noiseless case acting as the control scenario for our simulation. We see that both error metrics drop rapidly with increasing numbers of iteration and the final R-factor is approximately 1×10^{-3} indicating a perfect reconstruction; as to be expected for a case without noise. This also signals the correct functionality of the algorithm.

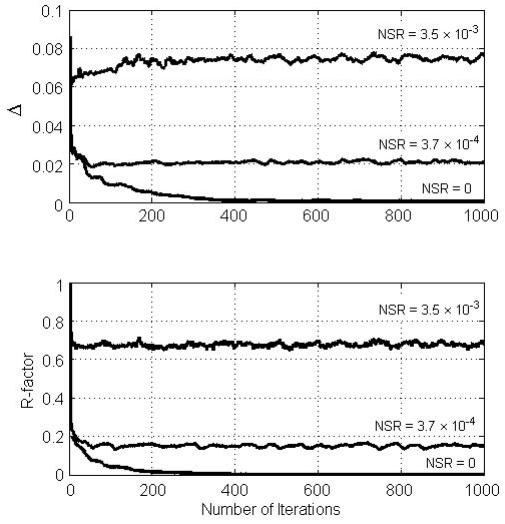
With a slightly elevated noise level, the reconstructed image is shown in the bottom right of the same figure. The noise-to-signal-ratio (NSR) for this case is around 3.7×10^{-4}, however after amplification, the NSR increases to around 4%. The reconstruction is still fairly reasonable as indicated by the final R-factor value of 0.14.

Lastly, a case where the noise level becomes too high for a good reconstruction to take place is shown in the bottom left. The original NSR before amplification is 3.5×10^{-3} whereas the NSR after amplification through the inverse averaged shape function is around 30% - quite a tough case to handle even for the DM algorithm. The final R-factor is 0.68. In figure 6, the mean crystallite sizes are plotted against the amplification of the energy of the Gaussian noise, i.e. the sum of the squares of all the random values used to corrupt the true magnitudes. A quadratic increase in the noise amplification is seen as the mean size of the nanocrystals gets larger. Note that even for small crystals, the noise effect is still quite significant as a mean crystallite size of 10 gives around a hundred-fold increase in the noise level.

Plots of the overall effect of different crystallite sizes and varying levels of measurement noise are then generated as shown in figure 7. The mean crystallite size was simulated from 5 to 30 in steps of 0.5 and the standard deviation of the Gaussian noise was increased from 0 to 0.7 in steps of 0.005. The noise levels are converted to NSRs before displaying on the axes of the mesh and contour plots. Each pair of crystallite size and noise level was simulated from the same starting point of the DM algorithm i.e. the same random image was used for the initial starting phase, and all ran for 1000 iterations. The plots were optimised in terms of the R-factor, i.e. the minimum R-factor value was graphed and the iteration that this occurs on was recorded and the delta metric corresponding to that iteration number was plotted.



(a)



(b)

Fig. 5. Results for a nanocrystallography phase retrieval simulation. (a) shows the reconstructions obtained from retrieving the phases using the DM algorithm; where clockwise from top-left is the original image, and then the reconstructed image for a noiseless case, a moderately noisy case, and a highly noisy case. (b) shows the Δ and R-factor error metrics resulting from the reconstructions in (a).

A distinct region of a hyperbolic shape can be seen on the contour plots. This is the convergence region of our particular nanocrystallography phase retrieval problem. The large flat region of the R-factor indicates that the metric never gets below its initial value during the 1000 iterations that

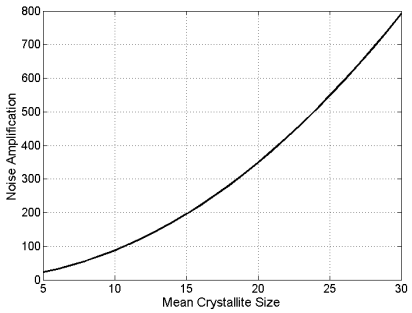


Fig. 6. Noise amplification with respect to the mean crystallite size in a collection of normally distributed nanocrystals with a standard deviation one-fifth of the mean.

the DM algorithm ran for and signals the breakdown in the convergence of phase retrieval from small crystals for this particular situation at hand.

VI. Conclusion

The diffraction from a nanocrystal corresponds to the molecular transform modulated by the transform of the crystallite shape function centred at the reciprocal lattice points. This means that the molecular transform can be measured between these lattice points, albeit the transform is attenuated; or equivalently, the noise is amplified. The result is that the amplitudes between the Bragg reflections are measured at a lower signal-to-noise ratio than for the amplitudes measured at the Bragg reflections themselves.

Simulation results on the performance of phase retrieval in the presence of a variable noise of this kind showed a hyperbolic convergence region when the mean crystallite sizes of the collection of nanocrystals were plotted against the measurement noise level.

The reconstruction was found to be fairly sensitive to noise and the success of phase retrieval was highly dependent on the noise level and crystallite size. Since the symmetries of the protein molecule used in our simulation have not been taken into account, incorporating this into our DM algorithm as another set of constraint might improve the overall sensitivity to measurement noise.

Other extensions of our current work include investigating the effect on the averaged shape function for crystallite sizes not normally distributed, and using crystals that are not square in shape. In addition, the simulated measurement noise was assumed to be Gaussian but that might not be true in practice where a proper statistical analysis is needed to determine the behaviour of the measurement noise on femtosecond X-ray detectors. These and other peculiarities of nanocrystallography await further investigation.

Ultimately it is hoped that macromolecules can be imaged effectively without the need for crystallisation and thus redefining and revolutionalising the field of X-ray crystallography and diffraction imaging.

Acknowledgment

The authors would like to thank Alok Mitra for providing the data for the AQP1 protein used in our simulations.

References

[1] P. Argos, "Protein crystallography in a molecular biophysics course," *American Journal of Physics*, vol. 45, pp.31-37, 1977.

[2] H. Chapman, "X-ray imaging beyond the limits," *Nature materials*, vol. 8, pp.299-301, April 2009.

[3] R. Neutze et al., "Potential for biomolecular imaging with femtosecond X-ray pulses," *Nature*, vol. 406, pp.752-757, 2000.

[4] D. Shapiro et al., "Powder diffraction from a continuous microjet of submicrometer protein crystals," *Journal of synchrotron radiation*, vol. 15, pp.593-599, 2008.

[5] R. Gerchberg and W. Saxton, "A practical algorithm for the determination of the phase from image and diffraction plane pictures," *Optik*, vol. 35, pp.237, 1972.

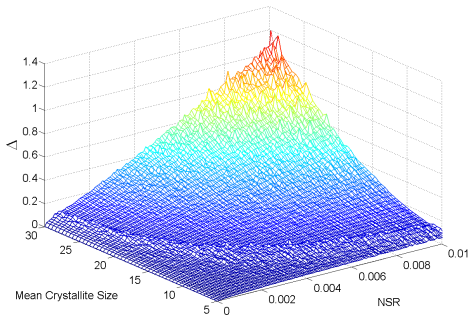
[6] J. Fienup, "Phase retrieval algorithms: a comparison," *Appl. Opt.*, vol. 21, no. 15, pp.2758-2769, 1982.

[7] S. Marchesini, "Invited Article: A unified evaluation of iterative projection algorithms for phase retrieval," *Rev. Sci. Instrum.*, vol. 78, no. 001301, 2007.

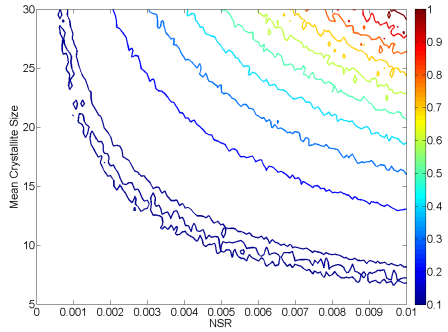
[8] V. Elser, "Phase retrieval by iterated projections," *J. Opt. Soc. Am. A*, vol. 20, no. 1, pp.40-55, 2003.

[9] A. Mitra et al., "Three-dimensional fold of the human AQP1 water channel determined at 4 A resolution by electron cyrstallography of two-dimensional crystals embedded in ice," *J. Mol. Biol.*, vol. 301, pp.369-387, 2000.

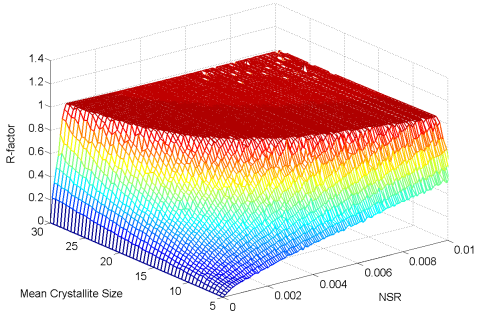
[10] D. Sayre, "Some implications of a theorem due to Shannon," *Acta Cryst.*, vol. 5, pp.843, 1952.



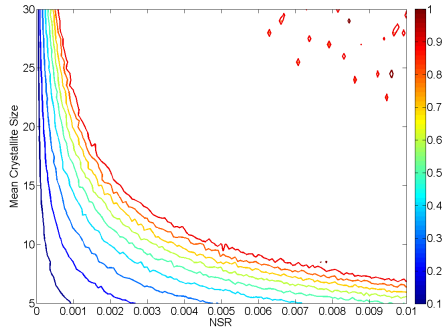
(a)



(b)



(c)



(d)

Fig. 7. Overall effect of varying the mean crystallite size and the noise level on the measured Fourier amplitude. (a) and (b) shows the mesh and contour plots respectively of the effect on the Δ metric of the DM algorithm, and (c) and (d) shows similar plots for the R-factor.