Phasing Two-Dimensional Crystal Diffraction Pattern with Iterative Projection Algorithms

by

Yun Zhao

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved September 2016 by the
Graduate Supervisory Committee:

John Spence, Chair
Nadia Zatsepin
Richard Kirian
Kevin Schmidt
Uwe Weierstall

ARIZONA STATE UNIVERSITY

December 2016

ABSTRACT

Phase problem has been long-standing in x-ray diffractive imaging. It is originated from the fact that only the amplitude of the scattered wave can be recorded by the detector, losing the phase information. The measurement of amplitude alone is insufficient to solve the structure. Therefore, phase retrieval is essential to structure determination with X-ray diffractive imaging. So far, many experimental as well as algorithmic approaches have been developed to address the phase problem. The experimental phasing methods, such as MAD, SAD etc, exploit the phase relation in vector space. They usually demand a lot of efforts to prepare the samples and require much more data. On the other hand, iterative phasing algorithms make use of the prior knowledge and various constraints in real and reciprocal space. In this thesis, new approaches to the problem of direct digital phasing of X-ray diffraction patterns from two-dimensional organic crystals were presented. The phase problem for Bragg diffraction from two-dimensional (2D) crystalline monolayer in transmission may be solved by imposing a compact support that sets the density to zero outside the monolayer. By iterating between the measured stucture factor magnitudes along reciprocal space rods (starting with random phases) and a density of the correct sign, the complex scattered amplitudes may be found (J. Struct Biol 144, 209 (2003)). However this one-dimensional support function fails to link the rod phases correctly unless a low-resolution real-space map is also available. Minimum prior information required for successful three-dimensional (3D) structure retrieval from a 2D crystal XFEL diffraction dataset were investigated, when using the HIO algorithm. This method provides an alternative way to phase 2D crystal dataset, with less dependence on the high quality model used in the molecular replacement method.

ACKNOWLEDGMENTS

Firstly, I would like to thank my advisor John Spence. He is very open-minded and high efficiency person with little formalism. And his passion on science motivates everyone around him. Besides, he is very knowledge in diffraction physics and have numerous brilliant ideas for my research projects. I feel extremely fortunate to work in our lab and pursue pure scientific goals without any distraction. Thanks a lot for his guidance on my project. And I'd also like to thank him for keeping me working on the same project even my progress went extremely slow for certain period.

Secondly, I would like to thank Nadia Zatzepin, Haiguang Liu, Shibom Basu who has helped me a lot when I first joined in this group. They taught me diffraction physics, data analysis and also guided my first project in Spence's lab. Thanks to Uwe Weierstall, Dingjie Wang, Dan James, who has taught me many experimental skills such as assembling nozzles, installing injector, operating jet etc. Thanks to Rick Kirian, Kevin Schmidt, Joe Chen for offering me suggestions on simulations.

I'd like to specially thank Chufeng Li who referred Spence lab to me. Chufeng Li, Ganesh Subramanian, Shibom Basu and Stella Lisova are my closest colleague and friends during my PhD. Thanks a lot for your advices, help and encouragement.

To my fellow PhD students Garrett Nelson, Jesse Coe, Chelsie Conrad and Shatabdi Roy Chowdhury, Gihan Ketawala, thanks a lot for your company and help. I enjoyed every minute when we were working together at SLAC beamtime. To our new members, Joe Chen, Natasha Stander and Andrew Shevchuk, thanks for your company and work.

Lastly, thanks to my family, mom, dad and my younger brother. Your love and support has sustained me through this whole hard time. I will forever be indebted.

TABLE OF CONTENTS

Page

APPENDIX Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

1.1 X-ray crystallography Overview

X-ray crystallography is a technique to determine the structure of crystals, in which periodically arranged atoms diffract X-ray to discrete called Bragg beams directions. The birth of this technology comes with the understanding of crystal properties as well as X-rays. Mankind has been admiring crystal's elegance for long time. The scientific study on crystallography started in 17th century when Johannes Kepler postulated that regular packing of water particles in snowflake rendered its hexagonal symmetry (Bencharit, 2012). In 1895, Wilhelm Roentgen discovered X-rays, at the time when the studies on crystal symmetry concluded (Assumus, 1995). In 1912, Max von Laue, inspired by Paul Ewald's doctoral thesis on crystal model, came up with an idea that the sub-micrometer spacing atoms in crystal might act as diffraction grating for X-rays [wiki]. Within the same year, Walter Friedrich and Paul Knipping conducted the first diffraction experiment on NaCl crystal as suggested by Laue [Figure 1]. Independently, W.L. Bragg and W. H. Bragg carried out similar experiments and provided the condition for finding a diffraction maxima with a very simple formula $2dsin(\theta) = n\lambda$ describing the relation among crystal lattice constant, incident X-ray wavelength and scattering angle, which is known as Bragg's Law. With the discovery of the mathematical formula, X-ray crystallography debuted modern science as an important probe for investigating structure of materials.

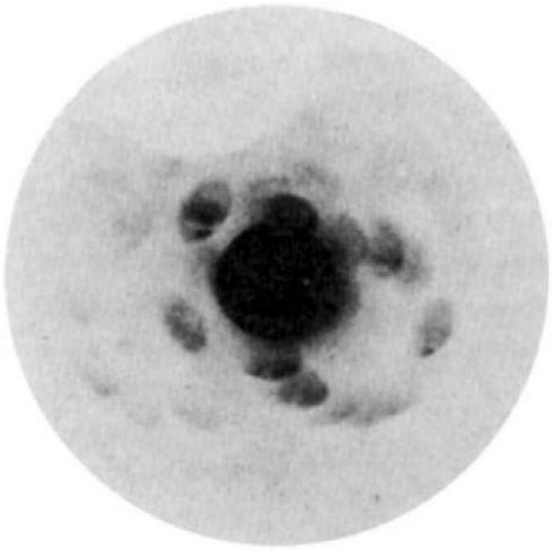


Figure 1.1 First diffraction pattern from NaCl crystals recorded by Walter Friedrich and Paul Knipping.

Since 1970s, the progress of science based on X-ray crystallography has been dramatic, largely due to the development of X-ray based technology, phasing theory and computation power (Hauptman, 1991). The advent of synchrotron radiation source improved X-ray beam brightness by ten thousand fold than previous lab-based sources. Besides, the implementation of charge coupled device (CCD) detectors in early 1990s further improved data collection speed and accuracy. The development of computer science enabled crystallography scientists to establish systematic methods to carry out most of the mathematically challenging work, including structure refinement and graphics computer-based model building. Molecular replacement, as an example, proposed by Rossmann (M. Rossmann, 1990; M. G. Rossmann & Blow, 1962), was a major breakthrough in bypassing the phase problem. As a result, X-ray crystallography has become one of the most commonly used techniques ever developed for the study of biomolecules at high resolution. More than 85% protein structures deposited in the PDB are solved by X-ray crystallography.

Despite the tremendous success of protein structure discovery at synchrotron based X-ray sources, traditional X-ray crystallography is mainly limited by radiation damage and sample preparation(Spence, Weierstall, & Chapman, 2012). Due to the presence of radiation damage, large crystals, at least micrometer in size, are required to sustain the radiation dose (or work around it). It may take years to find correct condition to grow large crystals that are suitable for diffraction.

The recent invention and development of the hard X-ray free-electron laser (XFEL) (R. a Kirian et al., 2010; Pellegrini & Stöhr, 2009; Schlichting & Miao, 2012; Spence et al., 2012) has opened up new opportunities for structural biology. Before the turn of the century, it was believed that true single-molecule imaging (Schlichting & Miao, 2012) using scattered radiation would never be possible because the radiation dose needed to achieve sufficient high-angle elastic scattering would, as a result of inelastic process, destroy the molecule. XFELs not only render diffraction data without radiation damage, but also gives alternative method for phasing(J. Miao, Kirz, & Sayre, 2000; J. Miao, Sayre, & Chapman, 1998; Jianwei Miao, Charalambous, Kirz, & Sayre, 1999).

1.2 X-ray free electron laser

Today, X-ray free electron laser, described as 4th generation photon sources, is the most advanced X-ray facility with performance exceeding the best of the 3rd generation storage rings based synchrotrons. Compared to other traditional sources, the XFEL features short intense pulses, fields of high amplitude and frequency and spatially coherence volume, which led to a genuine scientific revolution in X-ray crystallography. For example, a diffraction pattern can be recorded in about one second at synchrotron by exposing protein crystals to X-ray flux of about $10^{12} \sim 10^{13}$ photons/second(Hart et al., 2012). Because of the time duration, we only measure the average position of the vibrating atoms. XFELs, on the other hand, can produce the same amount of photons in femtoseconds, which enables us to take snapshots of molecular motion (atomic/lattice vibrations are typically in the 100s of fs to ps timescales) and make a molecular

movie. However, there is another concern: can crystals sustain such high beam power? Will radiation damage prevail in the XFEL experiment?

Radiation damage happens mostly in terms of ionization when a sample absorbs high doses of energy from incident X-rays. Apart from beam power, the frequency is also an important factor for ionization effect. The efficiency of absorption reaches a peak value when the electric field of the incident beam oscillates with approximately the same frequency as the orbiting valence electron, which is of the order of 10^{15}. For an incident beam with photon energy at 8 keV, its frequency is of the order of 10^{18}, which is 1000 times higher. Therefore, the high frequency has an effect to stabilize the atom against ionization. Simulation by Neutze et al showed that radiation damage can be outrun if the X-ray pulse duration is less than 50 fs, which is feasible with an XFEL.

The key physics behind XFEL is the self-organization phenomenon of electrons in a relativistic beam, in which an electron beam with random electron positions will change into a distribution with electrons regularly spaced at about the X-ray wavelength (Pellegrini & Stöhr, 2009; Schlichting & Miao, 2012). Typically, an XFEL consists of a linear accelerator followed by a long undulator magnet [Fig 2]. Bunch of emitted electrons from the source are first accelerated to several tens of GeV by a linear accelerator. When electron bunches moves into the undulator with a sinusoidal magnetic wave, they will follow the oscillating trajectory and emit electromagnetic radiation. The magnetic field not only changes the electron energy, but also modulates the electron beam to equal spacing bunches with the same period of radiation wavelength. Therefore, the electromagnetic waves produced by electrons superimpose in phase and result in a stronger field. In turn, the collective behavior of electrons become more effective. The net result is the exponential growth in the amplitude of electromagnetic wave and a fully coherent radiation emanating from the electron bunches. Hence the radiation intensity will be proportional to the square of number of electrons N_e^2. We should also note that, in storage ring based synchrotrons, this amplification factor is N_e as there is no correlation between electron positions on the scale of radiation wavelength.

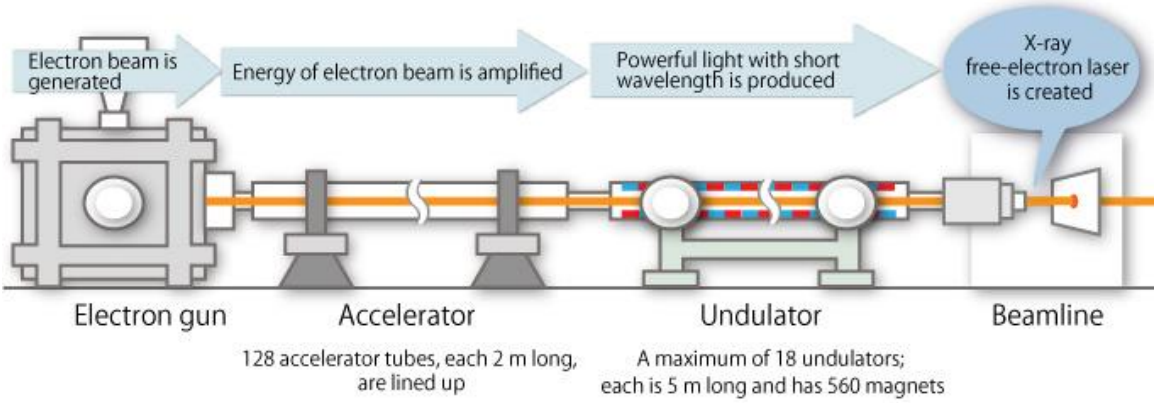


Fig 1.2 Schematic representation of a Free Electron Laser (Narumi & Sautter, 2011)

Currently, there are four XFEL facilities available for user experiments around the world. The Free electron Laser in Hamburg (FLASH) (M. J. Bogan et al., 2010; Michael J. Bogan et al., 2008) is the earliest soft XFEL source in operation from 2005, covering wavelength from 4.5 nm to about 47 nm with gigawatt peak power and 10~100 fs pulse duration. The first hard X-ray XFEL for experiments was the Linac Coherent Light Source (LCLS) at SLAC National Accelerator Laboratory, producing X-ray energy up to 9 KeV (wavelength 0.14 nm) with 3 mJ per pulse. The SPring-8 Angstrom Compact free electron Laser (SACLA) at the RIKEN Harima Institute in Japan (Chapman et al., 2011) and PAL-XFEL at South Korean started to operate in 2011 and 2015 respectively. Besides, more hard XFELs are under construction worldwide, including the European XFEL, Hamburg and the SwissFEL at the Paul Scherrer Institute, Switzerland (Schlichting & Miao, 2012).

1.3 Sample delivery at XFEL

In conventional X-ray crystallography experiments, many diffraction patterns can be collected from a single macroscopic crystal because the power of X-ray beam is relatively low. By gradually rotating a goniometer stage that holds the crystal, the orientation of the successive diffraction patterns can be recorded during measurement (Spence et al., 2012). At XFEL, X-ray pulses are so intensive that crystals will be destroyed once being hit. Instead of constantly shining X-rays on a crystal in synchrotron, an XFEL produces very short pulses, with a repetition rate of

120 HZ and 10~300 femtosecond pulse duration. As the pulse is so brief, the diffraction pattern recorded is actually from the intact structure before radiation damage takes place. As a result, the crystal can tolerate a significant higher dose than that at synchrotron. To fully take advantage of these features, developments on new sample delivery method as well as data analysis routine are demanded. Currently, there are three main forms of sample injectors designed for SFX experiments: the aerosol gas phase injector (M. J. Bogan et al., 2010; Michael J. Bogan et al., 2008; R. A. Kirian et al., 2015), the gas dynamic virtual nozzle (GDVN) liquid injector (U Weierstall, Spence, & Doak, 2012) and the lipid cubic phase (LCP) injector (Uwe Weierstall et al., 2014). Besides, a sample handling method is also developed by scanning fixed target, which has a potential for high hit rate (Hunter et al., 2014).

Aerosol injector

The aerosol injector was initially designed to deliver nanoscale particles for serial femtosecond X-ray diffraction experiments at FLASH. In this scheme, the sample of nanoparticles are generated using a charge-reduced nanoelectrospray aerosol source. Then a stack of aerodynamic lens are employed to focus aerosol particles into a stream of about 20 ~200 um in diameter at the point of intersection with the XFEL X-ray beam (Michael J Bogan, Starodub, Hampton, & Sierra, 2010). Hit rates from aerosol injector at the LCLS have increased from much less than 1% (early work) to about 10% on average, with a maximum 40%. The main advantage of an aerosol injector over liquid injector is the absence of background scattering from water jet in single particle X-ray diffractive imaging. Many types of aerosol sources can produce particles with unique size distributions. However, it may only apply for nanoscale materials such as core-shell structured atomic clusters, not for biomolecules (Michael J. Bogan et al., 2008).

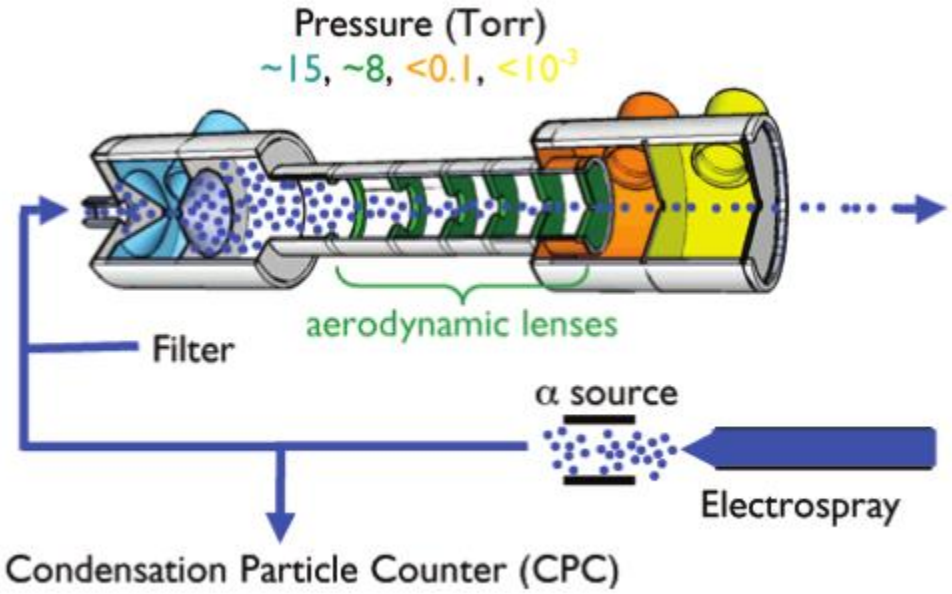


Fig 1.3 Schematic diagram of aerosol injector

GDVN

The GDVN liquid injection system, originally developed at ASU, delivers sample in a hydrated environment that is beneficial to preserve the native structure and function. This type of injector has been widely used for sample delivery in experiments such as protein solutions for wide angle scattering (Arnlund et al., 2014), nanocrystal suspensions for pump-probe time resolved crystallography (Aquila et al., 2012; Kupitz et al., 2014). The injection system consists a gas dynamic virtual nozzle and a long nozzle shroud. A scheme of the gas dynamic virtual nozzle is shown in Fig 1.4. The glass capillary at the center of nozzle carries the sample solution. Its tip is grained to a cone shape. Helium gas, which flows in between the glass capillary and glass tube, focuses the liquid to a straight line. The flow of gas and liquid are both driven by external pressure that can be controlled remotely through an HPLC or a gas regulator.

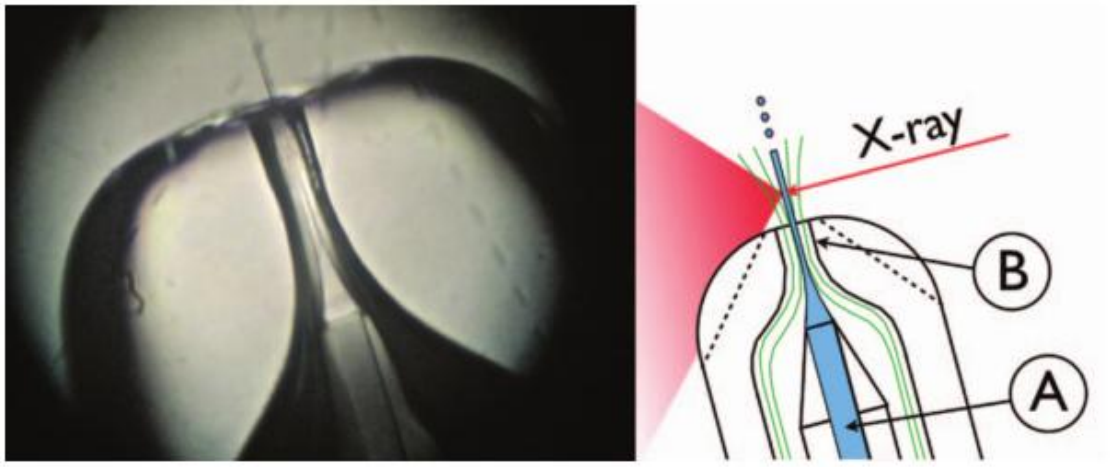


Figure 1.4 Gas dynamic virtual nozzle in operation and schematic(U Weierstall et al., 2012)

A straight and proper jet may be formed under proper pressure on the gas line and liquid line. Pressure is typically around 200~600 psi on gas line and 700~2000 psi on sample line. After the liquid flows out of the glass capillary, shearing gas focuses the jet to about 5 micron in diameter and accelerates its speed to about 10 m/s. The jet is operated at room temperature, typically at a flow rate 20 ul/min at CXI. Capillaries with 50 um, 75 um and 100 um ID were most often used, depending on crystal size and buffer condition. In order to save sample, low flow rates are preferred unless the jet disappears or breaks into droplets. Currently, the lowest flow rate achieved at the ASU lab is about 5 ul/min. The hit rate of a liquid injector mainly depends on the concentration of sample, stability of jet and beam position. Best case, the hit rate purely depends on the density of crystals as long as the jet is stable and X-ray beam hits the jet stream precisely.

Crooked jet and nozzle clogging are the two most common problems during XFEL experiments at CXI. Defects in nozzle parts, unbalanced pressure or liquid properties may cause the jet stream deflects away from central line. Practically, we only optimize the jet stream direction by trying out different pressure on gas line and liquid line when the sample is running with X-ray beam on. The defects from nozzles parts, such as asymmetric cone shape in the tip of glass capillary or gas aperture, can't be repaired or replaced within a reasonable amount of time even for an experienced nozzle technician. The clogging mostly happens either at the filter after

the sample reservoir or the nozzle tip. A quick and steep rise in the HPLC (control panel) pressure in combination with no visible jet flow indicates that either the nozzle is clogged or that the reservoir has run out of sample. A microscope fixed on the shroud of injector can directly observe the clogging at the nozzle tip. In this case, nozzle can be cleaned by running water and recycling. If an in-line filter gets clogged, then simply replacing with a new filter will suffice. Lastly, testing and characterization of sample injection in advance (before the experiment at LCLS) can significantly reduce the amount of problems during sample delivery.

LCP injector

The LCP injector was also originally developed at ASU. The design and principle of LCP injector are very similar to GDVN injector. The GDVN works well for fluids with low viscosity such as water. The main difference with the LCP injector lies in the pressure amplification design since a much larger pressure is required to inject a viscous jet. LCP offers advantages in that it can be used for both injection, and as a growth medium for membrane protein crystals (eg. G-protein coupled receptors (GPCR)) (Conrad et al., 2015; Liu, Wacker, Gati, Han, James, Wang, Nelson, Weierstall, Katritch, Barty, Zatsepin, Li, et al., 2013; Liu, Wacker, Gati, Han, James, Wang, Nelson, Weierstall, Katritch, Barty, Zatsepin, Li4, et al., 2013; Uwe Weierstall et al., 2014).

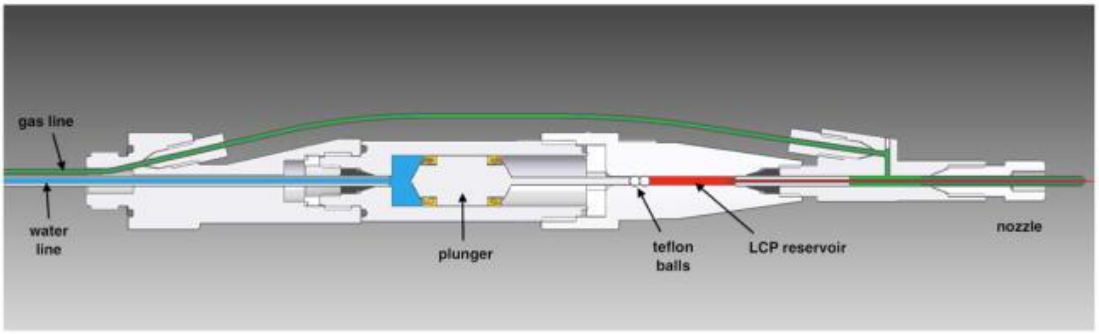


Fig 1.5 Middle section through the LCP injector

1.4 Data collection and analysis

XFEL detector

Many experiments at the LCLS require a detector that can image scattered X-rays on a shot-by-shot basis with high efficiency and excellent spatial resolution over a large solid angle

and both good S/N (for single-photon counting) and large dynamic range (required for the new coherent X-ray diffractive imaging technique). The Cornell-SLAC Pixel Array Detector (CSPAD) has been developed to meet these requirements. SLAC has built, installed, and characterized three full camera systems at the CXI hutches at LCLS (Hart et al., 2012).

Data analysis at XFEL

The data collected during an XFEL beamtime can result in 10-100 terabytes of data (transfer of data offsite may take many days). The first step in the analysis process, therefore, is data reduction. Data reduction is accomplished by software that finds frames where there are likely particle hits. The hit finding program Cheetah (Barty et al., 2014) is available freely under the GNU public license, and also provides useful online monitoring tools, that allow rapid feedback on data quality during the beamtime.

After data reduction (hit finding), particle orientation must be determined. Crystallographic indexing solves this problem for the SFX case. The data is then merged, phased (via known solutions to the crystallographic phase problem), and transformed to recover the electron density of the target molecule. Several software packages are now available for automating SFX data analysis (Sauter, Hattne, Grosse-Kunstleve, & Echols, 2013; White et al., 2013, 2012).

In terms of procedure, the Bragg peak positions and intensity values are firstly recorded. The crystal lattice type and lattice constant can be informed by measuring the angle and distance of Bragg spots. Then the Miller indices can be assigned to corresponding Bragg spots. Experimentally, the structure factor amplitudes are proportional to the square root of measured intensities of corresponding Bragg spots.

1.5 X-ray diffraction physics

X-rays mainly interact with electron cloud of the atom. So atoms with higher atomic number scatter X-ray more strongly. When X-rays reach an electron, several interactions may take place and emit secondary electromagnetic radiations (X-rays). According to the wavelength and phase relationship between incident wave and scattered wave, these interactions can be classified as elastic scattering, absorption, Compton scattering and fluorescence etc. In this

thesis, we will focus on elastic scattering, where the incident and outgoing electromagnetic wave have the same wavelength and phase over time and space. Another approximation made in the following introduction is that the scattering can be considered very weak so that multiple scattering events can be neglected. In this case, each diffraction pattern collected is the projection of a curved surface cut by the Ewald sphere in reciprocal space, which relates to the illuminated object by Fourier transform.

X-ray scattering by free electron

Free electron can be considered as the most elementary scattering unit in X-ray diffraction. The scattering of an X-ray by an electron can be perceived as follows. When an incident plane electromagnetic wave front hits an electron, the electron will oscillate under the force of the alternating electromagnetic field. The accelerating electron will act as another point source and radiate secondary spherical electromagnetic waves. The outgoing wave is given by

$$\overrightarrow{E_o}(R,t) = -\left(\frac{-e}{4\pi\varepsilon_0 c^2 R}\right)\vec{n}\times(\vec{n}\times\vec{a}(t')) \qquad (1.1)$$

where ε_0 is the permittivity of free space, c is the speed of light, R is the distance between electron and observation point, $\vec{n}$ is the radiation direction, t' is the retarded time, given by $t' = t - R/c$ and $\vec{a}(t')$ is the acceleration of the electron. For a linearly polarized incident wave with $\overrightarrow{E_i}e^{-i\omega t}$,

$$\vec{a}(t') = \frac{-e}{m_e}\overrightarrow{E_i}e^{-i\omega t'} \qquad (1.2)$$

where m_e is the mass of electron. Inserting equation (1.2) into (1.1), the magnitude of outgoing wave is

$$E_o(R,t) = -\frac{e^2 E_i}{4\pi\varepsilon_0 c^2 R m_e}\sin\alpha\, e^{-i\omega(t-\frac{R}{c})} = -\frac{r_e E_i}{R}\sin\alpha\, e^{-i\omega(t-\frac{R}{c})} \qquad (1.3)$$

where $r_e = \frac{e^2}{4\pi\varepsilon_0 c^2 m_e}$ is the classical electron radius. α is the angle between incident wave direction and outgoing wave direction. The time-averaged intensity of outgoing wave at R is

$$I_o = \langle|E_o(R,t)|^2\rangle_t = \frac{r_e^2}{R^2}I_i\sin^2\alpha \qquad (1.4)$$

For a pixel subtending a small solid angle $\Delta\Omega$, the collection area is $R^2\Delta\Omega$. So the photon intensity at that pixel is

$$I_o = \langle |E_o(R,t)|^2 \rangle_t = J_0 r_e^2 \sin^2 \alpha \, \Delta\Omega \qquad (1.5)$$

where J_0 is the incident photon flux density with unit number of photons/area.

Atomic Scattering factor

Atomic form factor describes the spatial intensity distribution of scattered X-ray by an isolated atom. An atom is composed of nucleus and electrons, both of which contributes to X-ray diffraction. However, the mass of nucleus is at least 10^3 times larger than electron. According to equation (1.2), the acceleration of nucleus is negligible compared with electrons. Therefore, the scattering effect of nucleus is often ignored in X-ray diffraction. When an atom with many electrons is exposed to a coherent incident X-ray beam, the outgoing electromagnetic wave is the coherent summation of all the outgoing waves from each electron at different positions, as shown in figure 1.6. For elastic scattering, the scattered wave preserves the same magnitude and phase of incident wave, while the propagation direction is changed. The scattering vector is defined as $\Delta\vec{k} = \overrightarrow{k_o} - \overrightarrow{k_i}$ and $\vec{q} = 2\pi\Delta\vec{k}$.

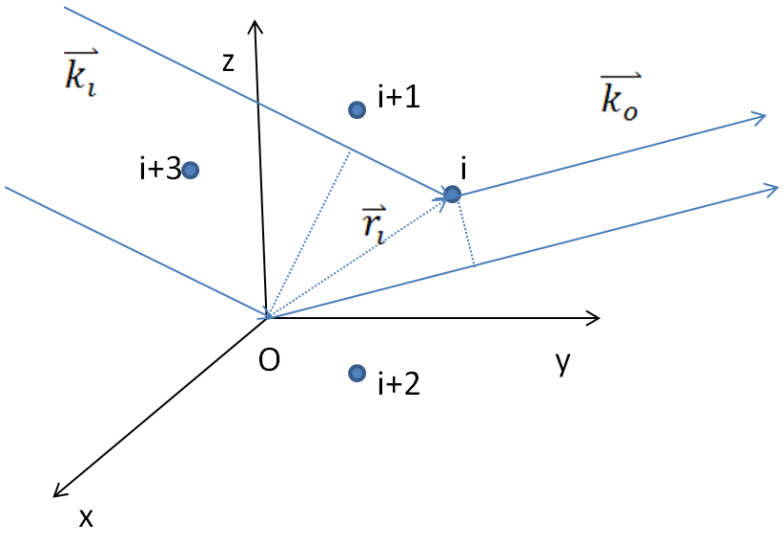


Figure 1.6 Scattering geometry from many electrons. $\overrightarrow{k_i}, \overrightarrow{k_o}$ are incident and outgoing wave vector respectively. $\vec{r_i}$ is the coordinate of the i-th electron in atom. The phase of electron i with reference to origin point O is $|\overrightarrow{k_i} * \vec{r_i}| + |\overrightarrow{k_o} * \vec{r_i}| = (\overrightarrow{k_o} - \overrightarrow{k_i}) * \vec{r_i}$.

The electron distribution of an atom is given by a probability distribution $\rho(\vec{r})$. Therefore, the atomic form factor can be expressed as

$$f(\vec{q}) = \int \rho(\vec{r})\, e^{i\vec{q}\vec{r}} d\vec{r} \tag{1.6}$$

A molecule is composed of atoms. Therefore, the scattering of a molecule is given by the sum of structure factors of each atom in molecule. The scattering factor of a molecule can be expressed as

$$F(\vec{q}) = \sum_i f_i(\vec{q}) * e^{i\vec{q}\vec{r}_i} \tag{1.7}$$

where $f_i(\vec{q})$ is the structure factor of the i-th atom.

The diffracted intensity from a molecule can be expressed as

$$I_o(\vec{q}) = \langle |E_o(R,t)|^2 \rangle_t = J_0 |F(\vec{q})|^2 r_e^{\,2} \sin^2 \alpha \, \Delta\Omega \tag{1.8}$$

X-ray diffraction from three-dimensional crystal

Now let's consider the X-ray diffraction from crystal. Let us assume that the structure factor of a unit cell with cell constant a, b *and* c is $F(\vec{q})$, where $\boldsymbol{q} = h'\boldsymbol{a}^* + k'\boldsymbol{b}^* + l'\boldsymbol{c}^*$. Here h', k', l' are fractional numbers and $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$ are called reciprocal space unit vectors, as shown in figure. They are called reciprocal because mathematically $\boldsymbol{a}^* * \boldsymbol{a} = \boldsymbol{1}, \boldsymbol{b}^* * \boldsymbol{b} = \boldsymbol{1}$ *and* $\boldsymbol{c}^* * \boldsymbol{c} = \boldsymbol{1}$.

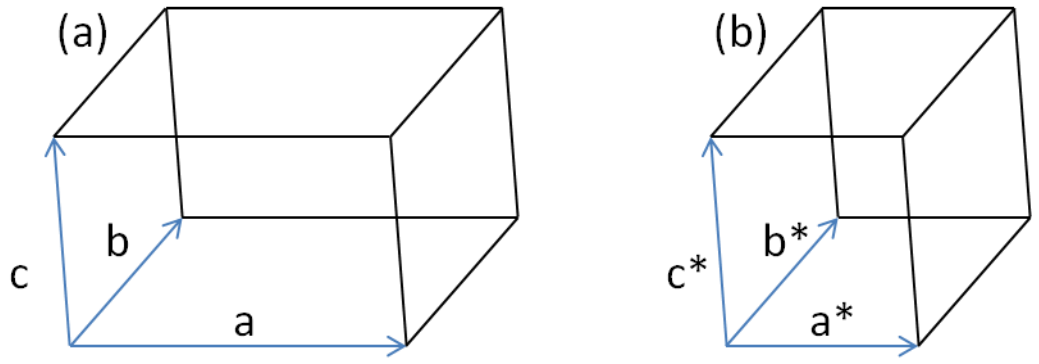


Figure 1.7 Real and reciprocal space

Then the scattering factor of unit cell with lower left corner at position $\boldsymbol{r} = n_1\boldsymbol{a} + n_2\boldsymbol{b} + n_3\boldsymbol{c}$ is $F(\boldsymbol{q}) * \exp(i\boldsymbol{q}\boldsymbol{r})$, where a, b, c is the lattice constant of unit cell. The scattering factor of the entire crystal is the sum of contribution from all unit cells.

$$F_{crystal}(\boldsymbol{q}) = \sum_{n_1,n_2,n_3} F(\boldsymbol{q}) * \exp\left[2\pi i\left(n_1 h' + n_2 k' + n_3 l'\right)\right]$$

$$= F(\boldsymbol{q}) \sum_{n_1,n_2,n_3} \exp\left[2\pi i\left(n_1 h' + n_2 k' + n_3 l'\right)\right] \quad (1.9)$$

$$= F(\boldsymbol{q}) * \sum_{n_1=0}^{N_1-1} \exp(2\pi i\, n_1 h') * \sum_{n_2=0}^{N_2-1} \exp(2\pi i\, n_2 k') * \sum_{n_3=0}^{N_3-1} \exp(2\pi i\, n_3 l')$$

Now let's examine the first summation term.

$$\sum_{n_1=0}^{N_1-1} \exp(2\pi i\, n_1 h') = \frac{1 - \exp(2\pi i N_1 h')}{1 - \exp(2\pi i h')}$$

$$= \frac{\exp(\pi i N_1 h')}{\exp(\pi i h')} * \frac{\exp(-\pi i N_1 h') - \exp(\pi i N_1 h')}{\exp(-\pi i h') - \exp(\pi i h')}$$

$$= \exp\{\pi i (N_1 - 1) h'\} * \frac{-2\sin(\pi N_1 h')}{-2\sin(\pi h')}$$

$$= \frac{\sin(\pi N_1 h')}{\sin(\pi h')} \exp\{\pi i (N_1 - 1) h'\} \quad (1.10)$$

Similarly, we obtain

$$\sum_{n_2=0}^{N_2-1} \exp(2\pi i\, n_2 k') = \frac{\sin(\pi N_2 k')}{\sin(\pi k')} \exp\{\pi i (N_2 - 1) k'\}$$

$$\sum_{n_3=0}^{N_3-1} \exp(2\pi i\, n_3 l') = \frac{\sin(\pi N_3 l')}{\sin(\pi l')} \exp\{\pi i (N_1 - 1) l'\}$$

The scattering factor of the entire crystal can, thus, be expressed as

$$F_{crystal}(\boldsymbol{q}) = F(\boldsymbol{q}) * \frac{\sin(\pi N_1 h')}{\sin(\pi h')} * \frac{\sin(\pi N_2 k')}{\sin(\pi k')} * \frac{\sin(\pi N_3 l')}{\sin(\pi l')} * \exp\{\pi i (N_1 - 1) h' + \pi i (N_2 - 1) k' + \pi i (N_1 - 1) l'\} \quad (1.11)$$

As mentioned in the previous section, the intensity distribution of a diffraction pattern recorded by a detector $I(\boldsymbol{q})$ is proportional to the square of the scattering factor modulus.

$$I(\boldsymbol{q}) \propto \left|F_{crystal}(\boldsymbol{q})\right|^2 = |F(\boldsymbol{q})|^2 * \left|\frac{\sin(\pi N_1 h')}{\sin(\pi h')}\right|^2 * \left|\frac{\sin(\pi N_2 k')}{\sin(\pi k')}\right|^2 * \left|\frac{\sin(\pi N_3 l')}{\sin(\pi l')}\right|^2 \quad (1.12)$$

Let's use $F_{shape}(\boldsymbol{q}, \boldsymbol{S})$ to replace the trigonometric terms

$$F_{shape}(\boldsymbol{q}, \boldsymbol{S}) = \left|\frac{\sin(\pi N_1 h')}{\sin(\pi h')}\right|^2 * \left|\frac{\sin(\pi N_2 k')}{\sin(\pi k')}\right|^2 * \left|\frac{\sin(\pi N_3 l')}{\sin(\pi l')}\right|^2 \quad (1.13)$$

where $\boldsymbol{S} = (N_1, N_2, N_3)$ and N_1, N_2, N_3 are the number of unit cell along each dimension of crystal. The term $F_{shape}(\boldsymbol{q}, \boldsymbol{S})$ is typically called shape transform because it depends on the shape and the size of crystals. The following figure shows the lattice grating interference function $\sin(N\pi x)/\sin(\pi x)$ for N =5,10 and 100.

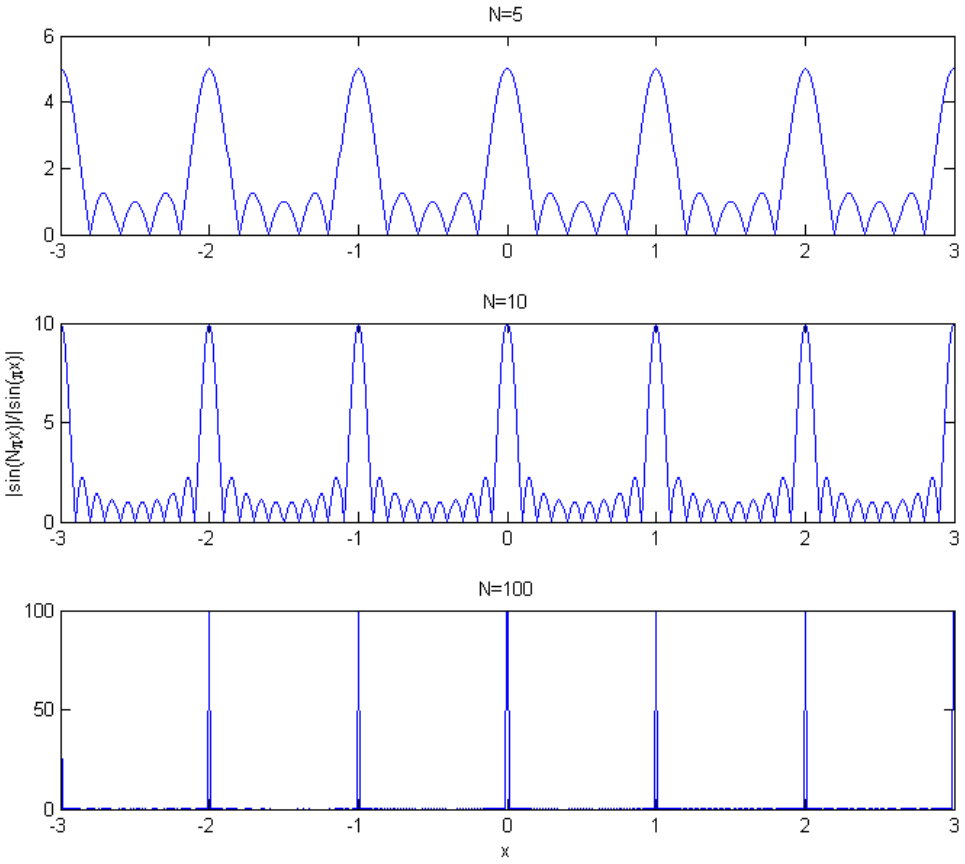


Figure 1.8 Lattice grating interference function.

Mathematically, it is easy to demonstrate that it has the following property

$$\left|\frac{\sin(N\pi x)}{\sin(\pi x)}\right| = \begin{cases} N\ (maxima), when\ x = 0, \pm 1, \pm 2, \dots \\ 0\ (minima), when\ x = \pm\frac{1}{N}, \pm\frac{2}{N}, \dots \end{cases} \quad (1.14)$$

In between the two adjacent major maxima, there are (N-1) minima and (N-1) secondary maxima that are smaller than the major maxima. The difference between major and secondary maxima will grow larger for increasing N. Therefore, in a nano-crystal where the number of repeating units in crystal is not very high, fringes can be observed in between Bragg spots (Chapman et al., 2011) [as shown in figure 1.9].

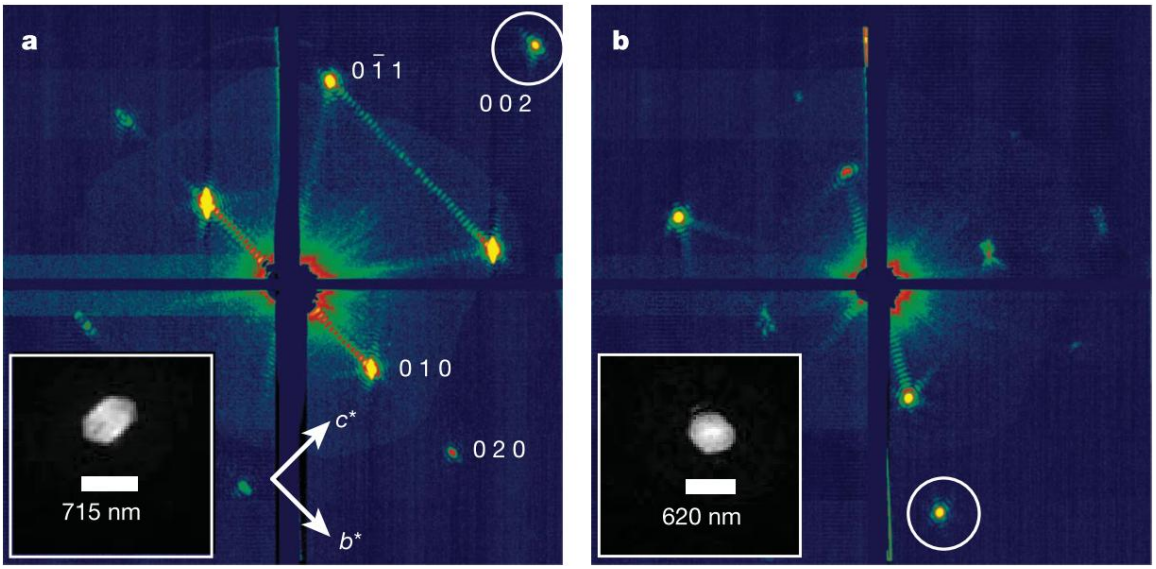


Figure 1.9 Shape transform from nanocrystal (Chapman et al., 2011).

For an infinite perfect crystal with $N_1, N_2, N_3 \to \infty$, the magnitude of major maxima will become dominant over secondary maxima and the appearance of lattice grating interference term approximates to Dirac comb functions (as shown in figure 1.8).

$$I(\boldsymbol{q}) \propto \begin{cases} |F(\boldsymbol{q})|^2 * {N_1}^2 * {N_2}^2 * {N_3}^2 & when\ h', k', l'\ are\ all\ integers \\ 0 & else \end{cases} \quad (1.15)$$

In this scenario, we can only observe sharp peaks at $\boldsymbol{q} = (h', k', l')$ with integer values. Recall that $\boldsymbol{q} = \Delta\boldsymbol{k}/2\pi$. This is exactly the Laue equation $\Delta\boldsymbol{k} * \boldsymbol{a} = 2\pi n_1;\ \Delta\boldsymbol{k} * \boldsymbol{b} = 2\pi n_2; \Delta\boldsymbol{k} * \boldsymbol{c} = 2\pi n_3$. The equation (1.15) also indicates that a larger crystal gives brighter and sharper Bragg spots.

In X-ray crystallography, the only direct data collected are the magnitude of structure factors $|F(\boldsymbol{q})|$, which is the Fourier transform of a single molecule. Using a crystal, instead of a single molecule, we may achieve a signal amplification of ${N_1}^2 * {N_2}^2 * {N_3}^2$ as indicated by (1.15). However, there is compromise. $|F(\boldsymbol{q})|$ is a continuous function over the full reciprocal space. But we can only measure intensities at the Bragg period from a crystal diffraction pattern, which under-samples reciprocal space by a factor of two. Therefore, we can't directly retrieve phase information using iterative projection algorithms, which is very successful for phasing single particle diffraction data. In sum, there is a trade-off between signal level and phase information in X-ray crystallography, when compared with single particle imaging.

X-ray diffraction from two-dimensional crystal

Membrane proteins can form natural two dimensional crystals (Pedrini et al., 2014). It can be considered as a special case of three-dimensional crystal with $N_3 = 1$, which means that there is only one layer along the z axis. Replacing $N_3 = 1$ to equation (1.12), the diffraction intensity from 2D crystal can be expressed as

$$I(\boldsymbol{q}) \propto \left|F_{crystal}(\boldsymbol{q})\right|^2 = |F(\boldsymbol{q})|^2 * \left|\frac{\sin(\pi N_1 h')}{\sin(\pi h')}\right|^2 * \left|\frac{\sin(\pi N_2 k')}{\sin(\pi k')}\right|^2$$

For an infinite and perfectly ordered 2D crystal with $N_1, N_2 \to \infty$, the lattice grating interference term approximates to Dirac comb functions (as shown in figure).

$$I(\boldsymbol{q}) \propto \begin{cases} |F(\boldsymbol{q})|^2 * {N_1}^2 * {N_2}^2 & when\ h', k'\ are\ all\ integers \\ 0 & else \end{cases} \qquad (1.15)$$

And its reciprocal space constitutes a set of rods perpendicular to the monolayer [as shown in figure].

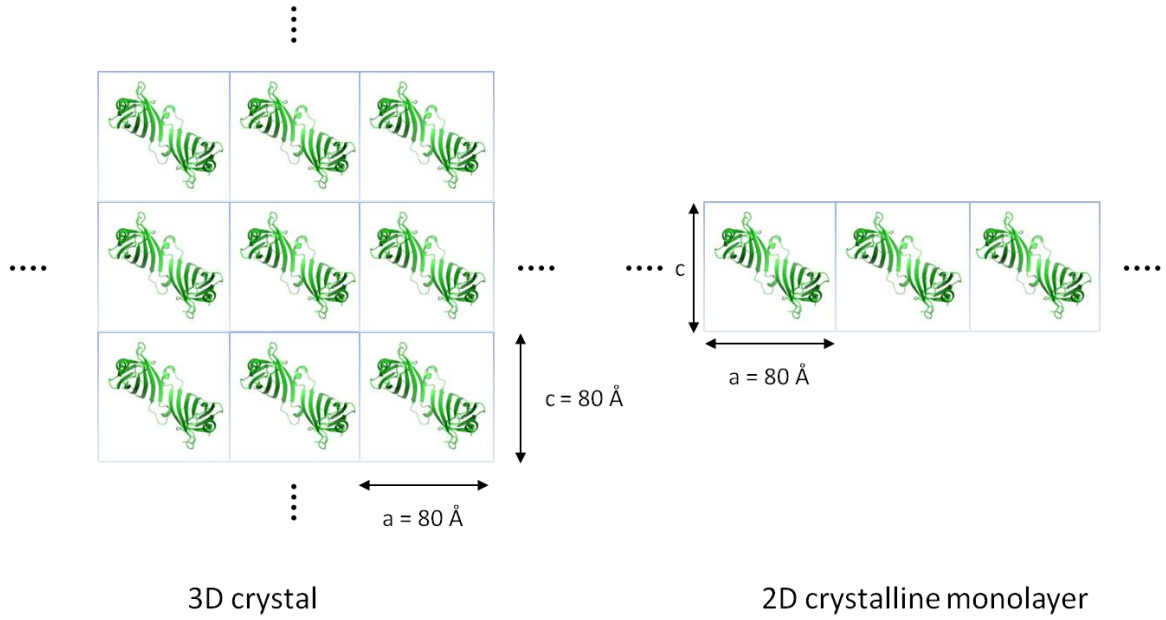


Figure 1.10 2D crystal monolayer. This figure shows the view along b axis direction.

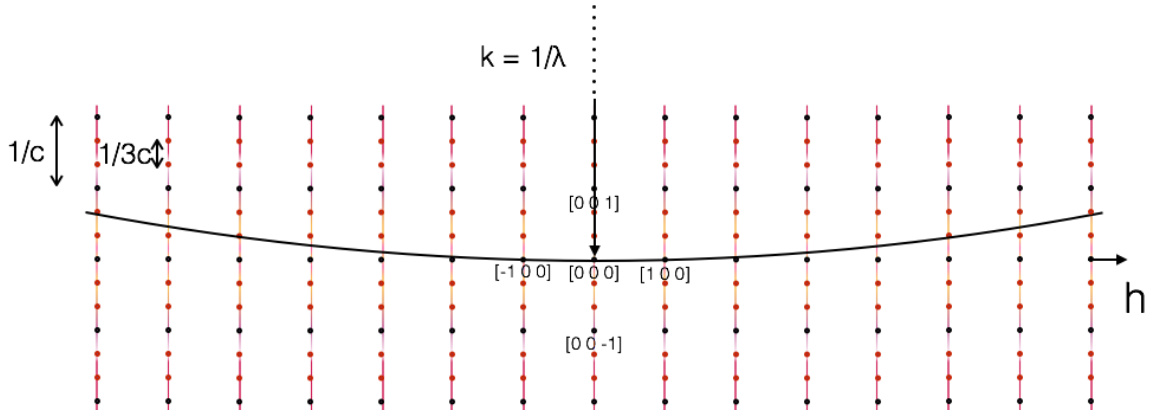


Figure 1.11 Reciprocal space of 2D crystal.

In comparison with a 3D crystal, a 2D crystal can be sampled as fine as we can along the z direction in reciprocal space, which may provide additional phase information. But the intensities in lateral direction are still under-sampled. In practice, the diffraction data set from 2-D crystal alone are typically insufficient enough to determine a unique structure.

1.6 Scope of this thesis

This thesis mainly discusses algorithms addressing image reconstruction and ab-initio phasing problems. Chapter 2 discusses the application of expectation and maximization algorithm in image reconstruction from extremely weak signals. Chapter 3 demonstrates the deconvolution of crystal powder diffraction patterns using auto-correlation algorithm. Chapter 4 introduces the phase problem and iterative algorithms for the case of two-dimensional crystals.

CHAPTER 2

STRUCTURE RECONSTRUCTION FROM EXTREME WEAK SIGNAL

2.1 Introduction

Much efforts have been devoted to study the structure of single particles with X-ray free electron laser (XFEL), which could produce very intense femtosecond X-ray pulse (Neutze, Wouts, van der Spoel, Weckert, & Hajdu, 2000; U Weierstall et al., 2012). This new method could potentially overcome the radiation damage on crystals as well as other limitations on traditional techniques (Fung, Shneerson, Saldin, & Ourmazd, 2008). However, each diffraction snapshot collected from a single particle contains very few photons, as the interaction between single particle and X-ray is too weak (Fung et al., 2008). An intuitive solution is through merging all the snapshots to obtain the complete diffraction pattern. The problem stems from the issue that we can't tell the orientation of particle just by each snapshot or by direct observation. Moreover, particles will be destroyed during each shot. The difficulty is exacerbated as the existence of background radiation noise. So a fundamental question in front of us is whether we are able to distinguish the orientation of two noisy diffraction patterns with sparse photons in principle.

One approach to classify the orientation is based on cross-correlation method by Huldt et al.(Hajdu, 2003). They successfully classified the diffraction patterns with approximately one photon per pixel. However, the photon fluence in our scenario is about 0.001 photons per pixel, much lower than Huldt's case. Hence, the cross-correlation method would fail in the ultra-low fluence limit (Philipp, Ayyer, Tate, Elser, & Gruner, 2012). Another robust method addressed to solving sparse randomly-oriented X-ray data was based on expectation-maximization(EM) method. EM method was first introduced to find parameters for a statistical model with incomplete data in information theory. Elser is one of the earliest people to have introduced this method in structure reconstruction from sparse randomly-oriented data (Elser, 2009; N.-T. D. Loh & Elser, 2009).

In this report, we focused on a 2D object with 4 random orientations during imaging, a much simpler case where I believe it is more illustrative to show the principle and feasibility of EM algorithm in structure reconstruction. So far, nobody has been able to reconstruct structure from

single snapshot with only one photon. Here we explore the minimum requirement for photon fluence to recover structure with given number of frames. Noise effects are also discussed. A detailed evaluation of EM algorithm for image reconstruction is also given the following parts.

2.2 Expectation and maximization algorithm

2.2.1 An intuitive explanation of EM algorithm

In general, EM method seeks to find some unknown parameters of a statistical model by iteration given measurement data, which contains some unobserved variables [8]. Below is an outline of EM iteration.

Let us assume a statistical model consisting of a set of observed data X, with missing values Z. We may start a random guess for unknown parameters θ. Then the likelihood function could be expressed as $L(\theta; X, Z) = p(X, Z|\theta)$. The maximum likelihood estimate of the unknown parameters is, then, determined by the marginal likelihood of the observed data

$$L(\theta; X) = p(X|\theta) = \sum_{Z} p(X, Z|\theta)$$

The iteration procedure is described as the following two steps [8]:

E-step: Estimate the expectation value of log-likelihood function, given distribution Z with parameter $\theta^{(i)}$ in ith iteration.

$$Q\left(\theta^{(i+1)}|\theta^{(i)}\right) = \langle \log\left(L\left(\theta^{(i)}; X\right)\right)\rangle = \langle \log\left(\sum_{Z} p(X, Z|\theta^{(i)})\right)\rangle$$

M-step: Determine the new $\theta^{(i+1)}$ which could maximize $Q\left(\theta^{(i+1)}|\theta^{(i)}\right)$.

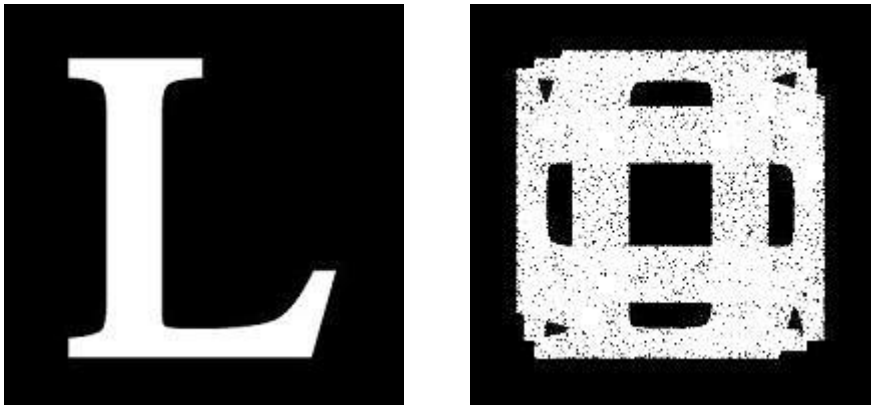
The parameter θ will converge to an optimal value by iteratively applying the above two steps.

One of the earliest paper on EM algorithm was by (Hartley, 1958). In that paper, he simplified the procedure for seeking the maximum likelihood computations of estimates from incomplete data by iteration. The iteration idea was also generalized to several cases. However,

the EM algorithm was first explicitly explained and given its name by a classic paper by Dampster, Laird and Rubin (Dempster, Laird, & B., 2007). They formulized the EM algorithm by defining expectation and maximization step with each iteration and generated its application to a wider class of statistical models. In particular, they also gave rigorous proof for the convergence of EM iterations for several models. More details on the convergence of EM algorithm can also be found in a book by G. McLachlan, and T. Krishnan (Mclachlan & Krishnan, 1977).

2.2.2 Data collection

The experiment designed here is almost the same as the one described in (Philipp et al., 2012). We simulated the imaging process of a 2D L-shape mask with extreme weak signals. The rotation of mask is spaced by $90°$. The detector in our simulation is a 200×200 pixel array. The orientation of mask will be reset randomly in one of the four equally possible orientations after an image is taken. Data sets with different quality are obtained by changing the photon counts per frame recorded during simulation. 10 000 snapshots were generated for each case.



(a) (b)

Fig 1. (a) The L-shape mask with a square aperture. (b) Sum of all frames with 40 photons per frame data set, showing a uniform distribution with 4 possible orientations.

2.2.3 Image reconstruction with EM algorithm

The algorithm we have adopted for the image reconstruction is based on the idea of expectation maximization. My interpretation here is largely based on several papers (Dempster et

al., 2007; Hartley, 1958; N. D. Loh et al., 2000; N.-T. D. Loh & Elser, 2009) and a book (Mclachlan & Krishnan, 1977). The derivation and idea are almost the same as Elser's work on reconstruction algorithm (N.-T. D. Loh & Elser, 2009; Philipp et al., 2012). Here I have presented more details and interpreted in a slightly different perspective, which perhaps easier to understand.

The parameter in the present setting is the intensity signal model w, a 200×200 matrix. The data collected are the sets of frames with photon counts k recorded by the detector, where the orientation of the mask relative to the detector r is intractable. Our model is updated, $\mathrm{w} \to \mathrm{w}'$, based on maximizing a log-likelihood function $\mathrm{Q}(\mathrm{w}')$. While orientation probability distribution of each frame $\mathrm{p_{rf}}$ is based on the current model parameters w. As we have 10 000 frames and 4 possible orientations, so $\mathrm{p_{rf}}$ is a $10\ 000 \times 4$ matrix in our algorithm.

Let's use $\mathrm{w_r}$ to denote the intensity distribution on detector when the image is in rotation r. The fth snapshot is assigned a probability distribution, $\mathrm{p_{rf}}$, with respect to its unknown rotation, r, relative to the current intensity model. The rotations are sampled in increments of $2\pi/\mathrm{N}$, where N defines the angular resolution of the reconstruction. N is 4 in our case as we know the number of possible orientations in imaging process in advance. Each frame comprises photon occupancy, $\mathrm{k_{if}}$, at pixel i, which in our low-fluence experiment are almost zero, the exceptions being equal to 1. Because the photon counts are independent Poisson samples of the intensity at each pixel, the probability is

$$\mathrm{p_{rf}} \propto \prod_{\mathrm{i}} \frac{\mathrm{w_{ir}}}{\mathrm{k_{if}!}} \mathrm{e}^{-\mathrm{w_{ir}}} \propto \prod_{\mathrm{i} \in \mathrm{I_f}} \mathrm{w_{ir}}$$

where ir is rotation r applied to pixel i, $\mathrm{I_f}$ is the set of pixels recording photons in frame f.

Then the probability of $\mathrm{f\_th}$ frame in orientation r could be normalized by

$$\mathrm{p_{rf}} = \frac{\prod_{\mathrm{i} \in \mathrm{I_f}} \mathrm{w_{ir}}}{\sum_{\mathrm{r}} \prod_{\mathrm{i} \in \mathrm{I_f}} \mathrm{w_{ir}}}$$

Note here that the probability is calculated by the current model w.

The log-likelihood function for f_th frame in orientation r is

$$Q_{rf}(w') = \log\left(\prod_i \frac{w_{ir}'^{\,k_{if}}}{k_{if}!} e^{-w_{ir}'}\right)$$

$$= \sum_{i=1}^{N_{pix}} \log\left(\frac{w_{ir}'^{\,k_{if}}}{k_{if}!} e^{-w_{ir}'}\right)$$

$$= \sum_{i=1}^{N_{pix}} (k_{if} \log w_{ir}' - w_{ir}' - \log k_{if}!)$$

As $k_{if} = 1 \text{ or } 0$, so $k_{if}! = 1 \text{ and } \log k_{if}! = 0$.

$$Q_{rf}(w') = \sum_{i=1}^{N_{pix}} (k_{if} \log w_{ir}' - w_{ir}')$$

Now the expectation of log-likelihood function may be written explicitly:

$$Q(w'|w) = \sum_{f=1}^{N_{frame}} \sum_{r=1}^{N_{rot}} P_{rf}(w) Q_{rf}(w')$$

$$= \sum_{f=1}^{N_{frame}} \sum_{r=1}^{N_{rot}} \sum_{i=1}^{N_{pix}} (P_{rf}(w) k_{if} \log w_{ir}' - P_{rf}(w) w_{ir}')$$

After obtaining the expectation estimate for $Q(w'|w)$, the algorithm proceeds to the second step.

w' is obtained by solving the equation $\frac{dQ(w')}{dw'} = 0$, as it should maximize the value of $Q(w')$. Note that $P_{rk}(w)$ comes from the expectation step, which depends on the current model w, rather than new model w'. So the maximizing update rule is given by

$$w_{ir} \to w_{ir}' = \frac{\sum_{f=1}^{N_{data}} P_{rf}(w) k_{if}}{\sum_{f=1}^{N_{data}} P_{rf}(w)}$$

Note that w_{ir} is the intensity of pixel i when the mask in orientation r. At the last step, we merge the models from different orientations.

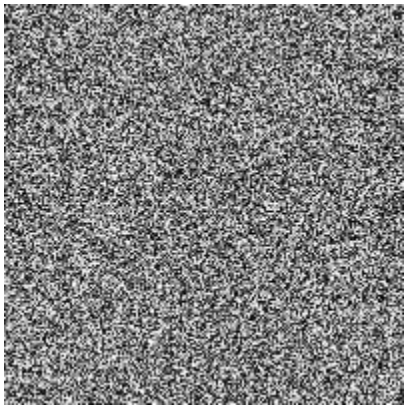
$$w'(i) = \langle \sum_{r} p_{rf} k_{-rf} \rangle_f$$

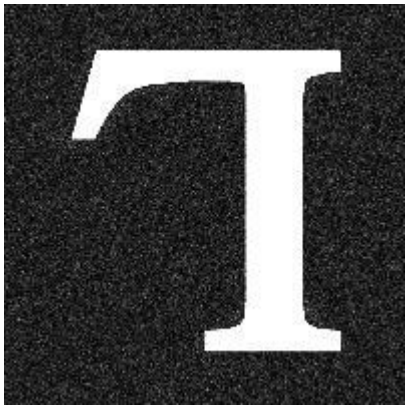
where $-rf$ means a rotation applied on frame f in the opposite direction of r.

The updated intensity model w' is an average of the photon counts in all frames with the appropriate distribution of rotations applied to each one. Each element in w' will be very tiny number after averaging, as each frame contains very few photons. In practical simulation, we need to amplify our final model w' by multiplying a proper constant to obtain a bright image, or it will be very dark.

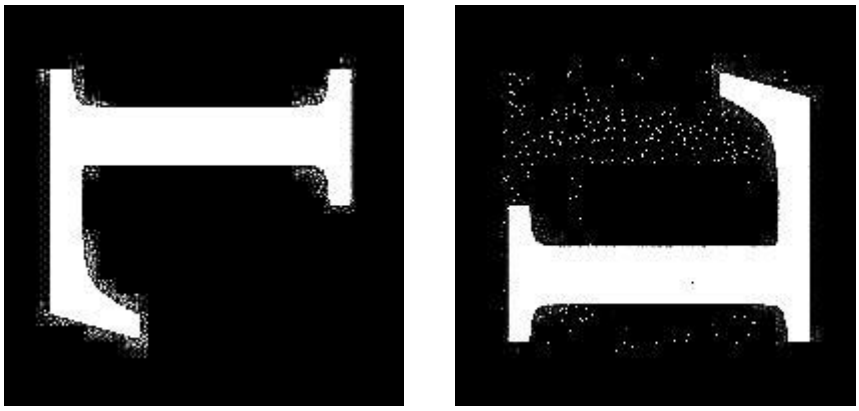
2.3 Image reconstruction

The EM iteration starts from a random model with each element assigned to a random number in the range of [0,1], as shown in Fig 2a. At the end of iteration, the model will end up a structure with arbitrary orientation. Figure 2a was reconstructed using 10 000 frames of data with an average of 40 photons per frame. This data set has a total of 0.5 million photons. For comparison, a data set with the same total frame but higher photon fluence was also processed. The reconstruction is shown in Fig 2d, where the average occupancy was 150 photons/frame.





(a) (b)



(c) (d)

Fig 2 (a) initial random model with no structural information. (b) A reconstruction using random-oriented data having a average 150 photons/frame with S/N=10; (c) A reconstruction using random-oriented data having a average 40 photons/frame; (d) A reconstruction using random-oriented data having a average 150 photons/frame.

The quality of the two reconstructions differ in classification accuracy, with the 150 photons/frames data yielding better results. There is also an increase in the iteration count of the EM algorithm: the 40 photons/frame data required 32 iterations, compared with only 4 iterations for the 150 photon/frame data. The minimum requirement for photon fluence is 40 photons/ frame, which is much higher than 2.5 photon/frame in Philipps' paper. This difference mainly comes from the fact that they have a much larger data set with 450 000 frames, which is 45 times bigger than here.

Images with noise are also studied here. We assume the background radiation is incoherent and uncorrelated between pixels. The net signal is simply the sum of X-ray scattering from mask as well as background. The S/N is defined as the average signal matrix element over the average noise matrix element. A successful reconstruction for an average 150 photons/frame with S/N=10 data set was shown in fig 2b. The presence of noise degrades the image quality and

raises the requirement for minimum photon fluence. The longest simulation made here is for 50 photons per frame data set with $S/N = 1$. It took for 287 iterations without any sign showing convergence.

In addressing noise problem, Elser gave the criteria for different classification methods (Elser, 2009). In that paper, he proposed that the arbitrarily high level of noise could be tolerated as long as unlimited measurements are available.

The EM algorithm demonstrated above could be generated to 3D reconstruction (N. D. Loh et al., 2000; N.-T. D. Loh & Elser, 2009). In that scenario, the 3D intensity model will be expanded into tomographic representation at first, as the information recorded by our detector is 2D information. This work was already done by Loh et al. and their code for a 3D particle reconstruction is available online (N. D. Loh, 2013).

The last comment we wish to make is about the limitation of the algorithm. The theoretical model matrix is pretty much binary as all the elements could just be 1 or 0, white or black in our image. Our approximation in p_{rf} estimation is greatly based on this assumption. If the elements in a model could be any real number between 0 and 1, can we still recover the model? The above algorithm failed to reconstruct it even with thousands photons per frames. A possible solution is that we just give up the approximation for Poisson distribution $\prod_i \frac{w_{ir}}{k_{if}!} e^{-w_{ir}} \propto \prod_{i \in I_f} w_{ir}$. But it will be computationally very expensive. In this regard, cross-correlation method seems to play a complementary role in addressing this problems.

2.4 Conclusion

The motif of this study was to demonstrate the principle of EM algorithm in sparse signal image reconstruction and classification. The minimum requirement for successful structure recovery depends on the photon fluence per frame, size of data set as well as S/N ratio. Comparing the simulation presented here with Philip's work, it seems that the minimum requirement for photon fluence can be relieved by producing a larger data set.

CHAPTER 3

DECONVOLUTION OF CRYSTAL POWDER DIFFRACTION PATTERN

3.1 Introduction

Rietveld refinement is a powerful approach to determine structure of crystals from powder diffraction data. Many programs available online have been developed based on this approach (Scardi, Mccusker, Dreele, Cox, & Loue, 1999). However, the success of this approach requires a good model first. In order to collect powder diffraction data, sample of small crystals have to be exposed to X-rays for long period of time, which may introduce significant radiation damage. Kam first pointed out that the three-dimensional structure of one particle may be determined using the X-ray scattering from many randomly oriented copies, without modeling of a priori information (Kam, 1977, 1980). Meanwhile, it was shown that the signal to noise ratio is the same for single particle and multiple particles per shot (R. a Kirian, Schmidt, Wang, Doak, & Spence, 2011). However, this method has remained undeveloped for about 20 years after Kam's paper due to the lack of brief and intense X-ray sources. With the availability of the free electron laser, this idea was re-discovered and the next stage of theoretical work is under development. Saldin et al performed many proof on principle simulations in single particle structure determination as well as experiments (Chapman et al., 2006; Saldin, Poon, Bogan, et al., 2011; Saldin & Shneerson, n.d.; Shapiro et al., 2008).

Here, we focus on the application of this method to crystal structure determination. Because the ensemble of crystals are static throughout the snapshot exposure, spinel crystals scattering patterns contain angular intensity fluctuations and thus differ from conventional powder diffraction pattern. These intensity fluctuations may provide us additional information on structure determination. It will be shown that the diffraction pattern for a single crystal can be recovered by fluctuation pair and triple correlation functions alone, without other apriori information.

3.2 Angular correlation function

3.2.1 *Spinel powder diffraction simulation*

For a coherent monochromatic plane wave, the incident and outgoing wavevector can be denoted as $\vec{k}_i$ and $\vec{k}_o$. The structure factor for a unit cell is given by

$$F\_cell(\vec{q}) = \sum_i f_i \exp(i * \vec{q} * \vec{r_i})$$

where $\vec{q} = \vec{k}_o - \vec{k}_i$, $\vec{r_i}$ is the atomic coordinates in unit cell, f_i is the corresponding atomic scattering factor.

The structure factor for lattice is given by

$$F\_lattice = \sum_n \exp(i * \vec{q} * \vec{r_n})$$

where $\vec{r_n}$ is the displacement of the nth unit cell with respect to origin. It will converge to a delta function as crystal becomes infinite.

Then, the scattering intensity from one crystal is

$$I(\vec{q}) \propto |F_{xtal}|^2 = \left| \sum_i f_i \exp(-\vec{q} * \vec{r_i}) \sum_n \exp(-\vec{q} * \vec{r_n}) \right|^2$$

Here we assume that different crystals scatter X-ray incoherently. Thus, the intensity observed on detector is simply the sum of the intensities from individual crystals.

$$I_k(\vec{q}) = \sum_i^{N_c} I(\vec{q}, \omega_k^i)$$

where ω_k^i is the orientation of i-th crystal during k-th snapshot. N_c is the number of crystals illuminated during k-th diffraction pattern. Because the number of crystals in correlated X-ray scattering is much less than in conventional powder diffraction, we may observe the spotty rings which reflect intensity fluctuations.

3.2.2 *Angular correlation function*

For the diffraction pattern of a single crystal, the pair correlation function for two different rings is defined as

$$C_1(q_i, q_j, \Delta\varphi) = \frac{1}{N_\varphi}\sum_{m}^{N_\varphi} I(q_i, \varphi_m) I(q_j, \varphi_m + \Delta\varphi) \qquad (1)$$

where q_i and q_j represents radius of the i-th and j-th ring on diffraction pattern.. N_φ is the number of azimuthal angels at φ_m which the intensity are measured. In a similar way, the triple correlation function is defined as

$$T_1(q_i, q_j, \Delta\varphi) = \frac{1}{N_\varphi}\sum_{m}^{N_\varphi} I(q_i, \varphi_m)^2 I(q_j, \varphi_m + \Delta\varphi) \qquad (2)$$

For many crystals case, the fluctuation pair correlation, which could be directly calculated from experimental data, is defined as

$$C_{exp} = \langle \left(\sum_{l}^{N_c} I(q_i, \omega_k^l) - < I_k(q_i) >_k \right) \left(\sum_{m}^{N_c} I(q_j, \omega_k^m) - < I_k(q_j) >_k \right) \rangle_k \qquad (3)$$

Then the pair correlation function for single crystal can be extracted by

$$C_1(q_i, q_j, \Delta\varphi) = \frac{1}{N_c} C_{exp} + \frac{1}{{N_c}^2} < I_k(q_i) >_k^2 \qquad (4)$$

In a similar fashion, the fluctuation triple correlation function is defined as

$$T_{exp} = \langle \left(\sum_{l}^{N_c} I(q_i, \omega_k^l) - < I_k(q_i) >_k \right)^2 \left(\sum_{m}^{N_c} I(q_j, \omega_k^m) - < I_k(q_j) >_k \right) \rangle_k \qquad (5)$$

Then the triple correlation function for single crystal can be extracted by

$$T_1(q_i, q_j, \Delta\varphi) = \frac{1}{N} T_{exp}(q_i, q_j) + \frac{2}{N} \langle I_k(q_j) \rangle_k C_1(q_i, q_j) - \frac{1}{N^3} \langle I_k(q_i) \rangle_k^2 \langle I_k(q_j) \rangle_k \qquad (6)$$

3.2.3 Reconstruction of single particle diffraction pattern

The intensity of a diffraction pattern can be expanded in circular harmonics as

$$I(q,\varphi) = \sum_m I_m(q)\exp(im\varphi) \qquad (7)$$

In general, $I_m(q)$ are complex numbers. Taking the Fourier transform of $C_1(q_i, q_j, \Delta\varphi)$ and $T_1(q_i, q_j, \Delta\varphi)$, we have

$$B_m(q_i, q_j) = \frac{1}{N_\varphi}\sum_{\Delta\varphi} C_1(q_i, q_j, \Delta\varphi)\exp(-im\Delta\varphi)$$

$$= I_m(q_i)I_m{}^*(q_j) \qquad (8)$$

$$FT_m{}^{(obs)}(q_i, q_j) = \frac{1}{N_\varphi}\sum_{\Delta\varphi} T_1(q_i, q_j, \Delta\varphi)\exp(-im\Delta\varphi) \qquad (9)$$

and it can be shown that

$$FT_m{}^{(calc)}(q_i, q_i) = I_m{}^*(q_i)\sum_{M\neq 0,m} I_M(q_i)I_{m-M}(q_i) \qquad \text{for } m \neq 0. \qquad (10)$$

So the magnitude of $I_m(q_i)$ is determined by $|I_m(q_i)| = \sqrt{B_m(q_i, q_i)}$. The unknown phases needs to be determined to reconstruct the single crystal diffraction pattern.

3.3 Application to spinel powder diffraction pattern

3.3.1 *spinel powder diffraction pattern simulation*

Each spinel crystal has 10 unit cells in x and y direction, with a lattice constant of 8.0858 Å. The wavelength of the incoming X-ray is 1.5406 Å. A flat Ewald sphere is assumed in the present simulation. The simulated diffraction pattern from single crystal is shown as follow

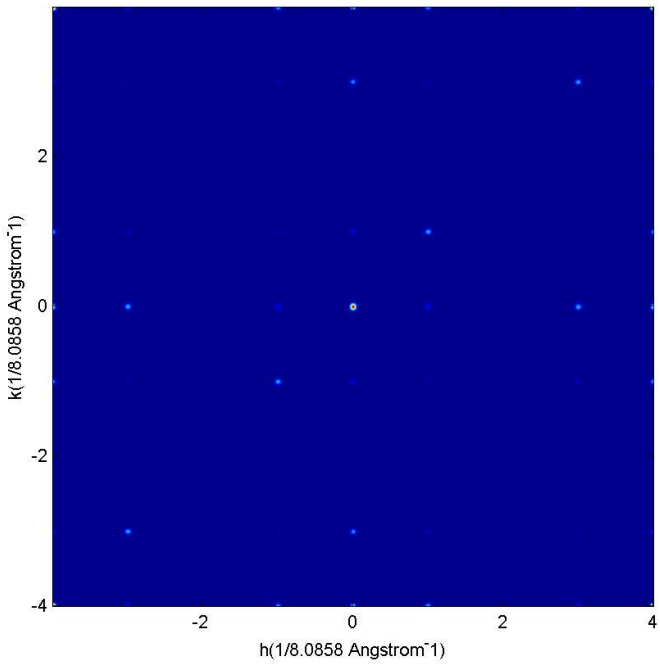


Figure 3.1 Diffraction pattern for single crystal

Next we simulate powder diffractions where 10 crystals are illuminated simultaneously per shot. Each crystal lies in a random orientation along z axis and scatters X-rays incoherently. In this way, we may obtain spotty powder diffraction rings, as shown in Fig 3.2.

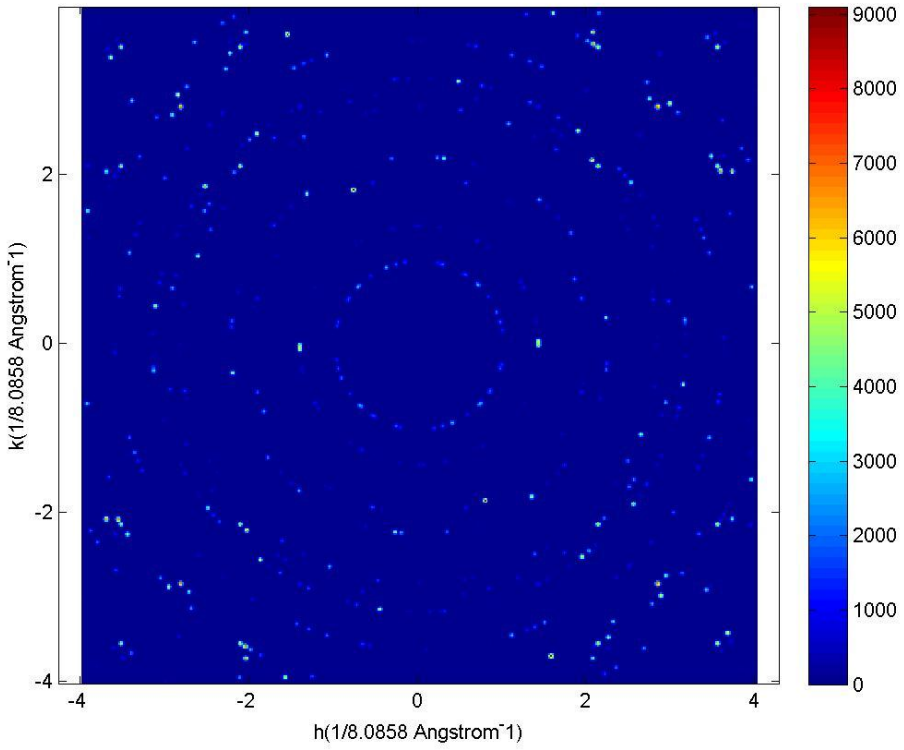


Figure 3.2 Diffraction pattern for 10 crystals with random orientation

In this report, we mainly investigate whether we can recover the diffraction pattern for a single crystal (Fig 1) from powder diffraction data (Fig 2). First, we need to obtain convergent values for angular correlation functions by averaging them over a large number of multiple-crystal diffraction patterns. In this case, 100 diffraction patterns were simulated. The averaged angular autocorrelation function shows the convergence to single crystal (Fig 3).

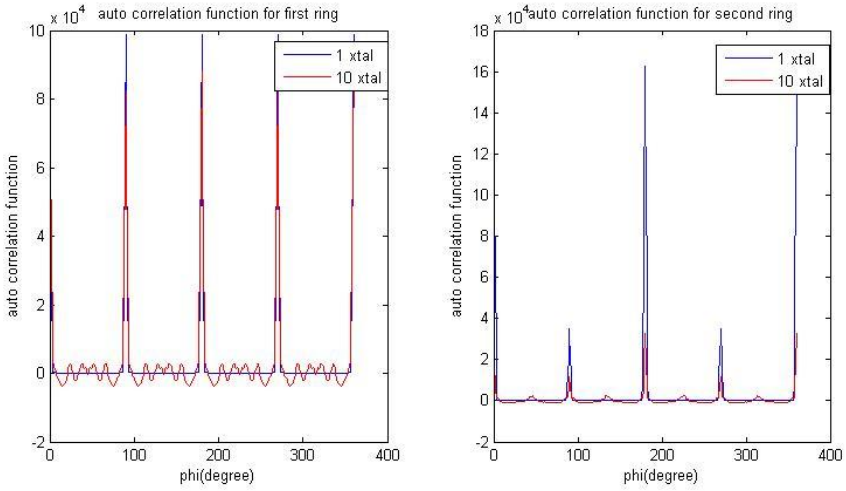


Figure 3.3 Auto correlation function retrieval for first and second ring

The magnitude of $I_m(q_i)$ can be uniquely determined by taking the square root of $B_m(q_i, q_i)$. Its phase could be solved by the charge-flipping method described in [8]. In present report, we take all $I_m(q_i)$ to be real and maximum value of m is 38. Note that $I_{-m}(q_i) = I_m{}^*(q_i)$ as a result of Friedel's rule. So only even values are non-zero. Here we take all coefficients as real. Only the parity (+/-) signs need to be determined. After searching 2^{19} combinations of signs to optimize the function.

$$\sum_{m \neq 0} \left| FT_m{}^{(obs)} - FT_m{}^{(calc)} \right|^2$$

The result of reconstructing the single diffraction pattern is shown in Fig 4.

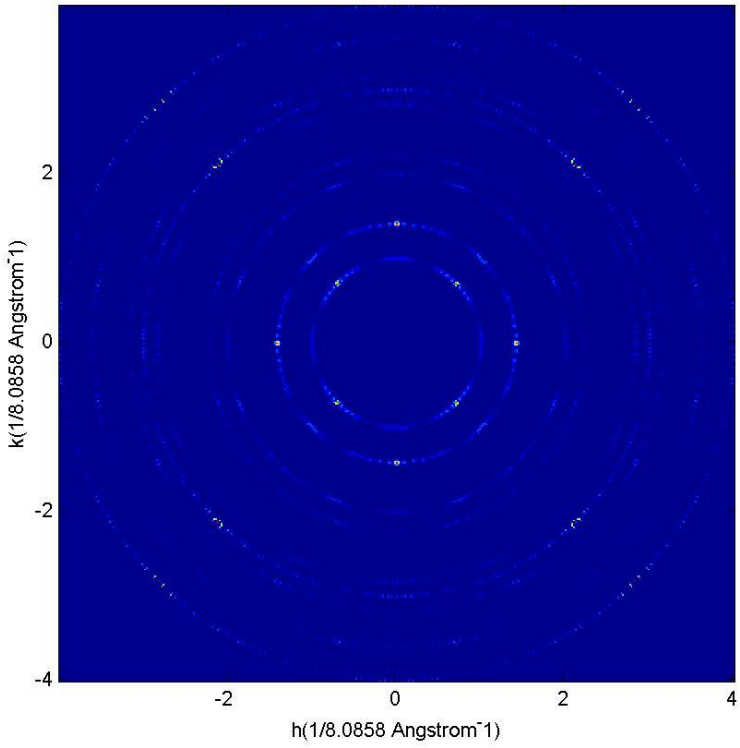


Fig 4 Single crystal diffraction pattern reconstructed from the magnitude of $I_m(q_i)$ determined from the mean pair correlation C_1, and signs from the mean triple correlations T_1 from 100 multi-particle diffraction pattern like that of figure 2.

3.3.2 *3D structure determination*

So far, we have reconstructed the 2D low-resolution diffraction pattern. More efforts are still required to develop this method to 3D reconstruction before real application to a powder diffraction experiment. Firstly, we cannot obtain a powder diffraction pattern just by rotating the crystal along one axis in a real experiment. All orientations need to be adequately sampled. Secondly, we should note that the diffraction pattern reconstructed is low-resolution data. For high-resolution data, the diffraction pattern will be the projection from curved Ewald sphere.

As yet, no simulation or experiment work on real 3D structure reconstruction has been successfully performed by this method. The low-resolution diffraction patterns probably originate

from the assumption of flat Ewald sphere. Elser generalized this method to a semi 3D case (Shapiro et al., 2008), where particles can be aligned in random orientations on a 2D substrate which can tilt freely with respect to the X-ray beam. As the tilt angle between substrates and X-ray beam can be measured and the correlation function has the same property as eqn (8) and (10), the reconstruction proceeds pretty much similar to the case for 2D case (Elser, 2011).

For the full rotation freedom case, the reconstruction idea is still the same. Firstly, we need to obtain convergent pair and triple correlation functions from simulated powder diffraction patterns. Then we expand the 3D reciprocal-space map by spherical harmonics (R. a Kirian, 2012).

$$I(\vec{q}) = \sum_{lm} I_{lm}(\vec{q}) Y_{lm}(\vec{q})$$

It can be shown that

$$C_1(q_i, q_j, \Delta\varphi) = \frac{1}{4\pi} \sum_{l}^{l_{max}} P_l(\cos[\Delta\phi]) B_l(q_i, q_j)$$

where $C_1(q_i, q_j, \Delta\varphi)$ is the ring cross correlation, $P_l(\cos[\Delta\phi])$ are the Legendre polynomials, and

$$B_l(q_i, q_j) = \sum_{m=-l}^{l} I_{lm}(q_i) I_{lm}^*(q_j)$$

Then we need to find all the complex coefficients involved from the above equation. It is a formidable task either using triple correlation method or phase iterative method (Saldin, Poon, Schwander, Uddin, & Schmidt, 2011).

3.4 Conclusion

Here we mainly demonstrate that the 2D diffraction pattern from single crystal can be reconstructed from powder diffraction data, in principle. There are still several limits on the present 2D simulation. First, we may observe that the intensity of (110) spot is not equivalent to $(\bar{1}10)$ from the single particle diffraction pattern. But the reconstructed diffraction pattern couldn't

distinguish this pair. From the pair correlation function, we could observe the intensity variation. It seems that this inefficiency doesn't originate from the expansion order m_{max}, but the accuracy of phase where all coefficients are assumed real. Secondly, a proper reference ring is crucial for successful reconstruction both in triple correlation method or phase iterative method. In this report, we chose the first ring as our reference ring and then calculated the pair correlation function with respect to the first ring, which indicates the relative position information of Bragg spots on different rings. The phases of high-resolution rings are not well recovered. It may be improved by choosing several outer rings as reference ring (Saldin et al., 2010).

Although the diffraction pattern reconstruction demonstration in this report is two dimensional, this idea provides us many insights on the application of real 3D powder diffraction. As to the 3D diffraction volume reconstruction, substantial research efforts are required to develop a functional theory.

CHAPTER 4

PHASING TWO-DIMENSIONAL CRYSTAL DATA WITH ITERATIVE PROJECTION ALGORITHM

4.1 Phase problem

In typical X-ray crystallography experiments, the major data are 2-D diffraction patterns produced from the X-rays scattered by a crystal. The routine data analysis can be performed using two steps, indexing and phasing respectively. In the indexing step, the amplitudes of complex structure factors $|F_{hkl}|$ can be calculated after mapping the Bragg spot intensities I_{hkl} back into 3-D reciprocal space. However, the associated phases φ_{hkl} cannot be measured directly from X-ray diffraction pattern alone. Therefore, the experimental information is intrinsically deficient for solving the 3-D structure, which constitutes the famous phase problem. Phase retrieval is a general problem based on assumptions. For example, we suppose that the object is finite, positive density etc.

Besides X-ray crystallography, the phase problem exists in many other fields as well, such as general X-ray diffraction, electron diffraction, neutron diffraction, astronomy etc, where only magnitudes of the Fourier transform of object density can be measured (J. Miao, Ishikawa, Robinson, & Murnane, 2015; Shechtman et al., 2015). Its importance can never be overstated. Currently, various phasing methods have been developed to address the phase problem for both periodic as well as non-periodic objects. For example, molecular replacement is the most widely used phasing method for protein crystallography. About 70% of the deposited structures in PDB are solved by molecular replacement. In the case of non-periodic objects, the Hybrid Input-Output algorithm is a very successful algorithm to solve the structure by iterating between real and Fourier space (Chapman et al., 2006; Jianwei Miao et al., 1999; Seibert et al., 2011).

4.1.1 *Phasing method in crystallography*

Crystals are often treated as infinite in crystallographic data analysis. The boundary of the molecule can hardly be estimated from Patterson function, unless the unit cell is almost empty. Therefore, real space information can hardly be inferred from external assumption, which is the

case for non-periodic objects. Today, molecular replacement (MR), first proposed by Micheal G Rossmann in 1962 (M. Rossmann, 1990), is the most popular method for crystallographers to get initial phases. In MR, the initial electron density map is estimated by performing inverse Fourier transform of complex structure factors, which combines experimental structure factor amplitudes with phases from model, which should be similar to our target structure. Actually, MR was firstly used as a phasing method for identical proteins crystallized in different space groups, mutant screenings or multiple ligand-target complexes. Because a large number of protein structures are readily available in the PDB (~100, 000), the probability of finding a reasonably good starting model for MR is quite high. Even partial search models can be successfully used for phasing with MR, making it a very powerful technique to obtain phases for crystallographic data.

Quite a few experimental phasing methods were developed before MR, such as multiple heavy atom isomorphous replacement (MIR) and single heavy atom isomorphous replacement, where the Bragg intensity differences between the heavy atom labeled crystals and the native crystal (Hendrickson, 2013) were compared. However its practical implementation is often difficult or time- and labor consuming. For small molecules, typically less than 1000 atoms per unit cell, this problem is usually addressed by applying direct methods, which solely use information from structure factor amplitudes and exploit chemical constraints to derive the phases of different Fourier components.

4.1.2 *Phasing methods for non-periodic object*

Hybrid input-output(HIO) algorithm is one of the most successful algorithms developed to address phase problem for a non-periodic object. The iterative algorithm imposes constraints between real space (support) and reciprocal space (structure factor amplitude) respectively. The support specifies the boundary of object. The density values outside of the support are declared to be zero. The first object density estimate is the inverse Fourier transform from known structure factor amplitudes and random phases. Then the density values outside of support are changed to zero. Then new phases are estimated by performing Fourier transform of modified object density. The next object density is calculated by doing inverse Fourier transform of phases from the

previous step and known structure factor amplitudes. The solution will be optimized after certain number of iterations.

The only prior information required for HIO algorithm is the support. There is a natural advantage for single particle diffraction. The autocorrelation function of object density can be obtained by doing inverse Fourier transform of intensities of diffraction pattern, which are proportional to the square of structure factor amplitudes. The autocorrelation is the twice of the object density in each dimension. For single particle diffraction, one implicit prior assumption is that the particle size is finite, so that it safe to claim that electron densities are zeros outside of certain boundary. If we know the size estimate, then the support can be a rectangular shape or box. Even if no size information available, the boundary can be estimated from the autocorrelation function which has a boundary with given finite object density.

4.1.3 *Uniqueness of phasing problem*

Before we apply any phasing methods to a diffraction dataset, it's very useful to examine the phase problem from a basic mathematical point view. Here we limit our discussion to the kinematic X-ray diffraction experiment, in which case the object density is the inverse Fourier transform of reciprocal space as shown in equation (4-1 & 4-2).

$$\rho(\vec{r}) = IFFT\big(F(\vec{q})\big) = \sum_{\vec{q}} F(\vec{q}) * \exp(-2\pi i * \vec{q} * \vec{r}) \qquad (4.1)$$

$$F(\vec{q}) = FFT\big(\rho(\vec{r})\big) = \sum_{\vec{r}} \rho(\vec{r}) * \exp(2\pi i * \vec{q} * \vec{r}) \qquad (4.2)$$

Solving the phase problem is equivalent to solving structure in real space. If we completely know the object density distribution $\rho(\vec{r})$ in real space, then we can calculate its complex structure factors $F(\vec{q})$ by equation (3.2). On the other hand, if we can measure both the amplitudes and phases of $F(\vec{q})$ through experiment, then we may solve the correct structure by equation (4.1). However, the only information we can extract from X-ray diffraction dataset is structure factor amplitudes. Therefore, prior knowledge is required to solve the correct density map. The prior knowledge can be any constraints in real space such as real and positive density, object size or

envelope, finite boundary, etc. Knowledge on low-resolution phases also serves as effective constraints to reduce the freedom of possible solutions.

It is often convenient, both for data processing and for the purpose of mathematical analysis, to represent a 2-D image or 3-D object by a discrete array of its sampled values. Intuitively, it is clear that if these sampled values are taken sufficiently close to each other, the sampled data are an accurate representation of the original function. Ideally, we need infinite number of infinitesimal pixels to accurately represent a continuous density distribution, which means an infinitely high resolution. In practice, we take sample values as long as it can accurately represent our object. First of all, any real experimental measurement has an upper resolution limited either by instrumentation, or sample quality etc. Our eyes have a limited resolution too. Most people can barely distinguish two points separated by 0.3 m that are 1 km away. Therefore, as long as the sampling interval in real space is fine enough for our purpose, there is no benefit in increasing the sampling resolution and collecting additional information. Secondly, more sample points also means bigger input 3-D array, which will take more memory and cause our program run for a much longer time.

The uniqueness of phase problem can be better illustrated by digitizing real as well as reciprocal space into a discrete numerical 3-D array. Then equation (4.1) can be reformulated into a set of linear equations. Assuming that we take sample values at equal spacing on object density $\rho(\vec{r})$ as well as structure factors $F(\vec{u})$, then $\rho(\vec{r})$ and $F(\vec{u})$ can be represented as discrete 3-D arrays ρ_{x_1,x_2,x_3} and F_{q_1,q_2,q_3} with size N_1 by N_2 by N_3. Here $x_1, x_2, x_3, q_1, q_2, q_3$ are integers and $x_1, q_1 = 0, .., N_1 - 1$; $x_2, q_2 = 0, \dots, N_2 - 1$; $x_3, q_3 = 0, .., N_3 - 1$. The total number of elements in each array is $N = N_1 \times N_2 \times N_3$. The equation (3.1) can be represented as discrete Fourier transform

$$\rho_{x_1,x_2,x_3} = \frac{1}{N}\sum_{q_1,q_2,q_3} F_{q_1,q_2,q_3} * \exp\left\{-2\pi i * \left(\frac{x_1 q_1}{N_1} + \frac{x_2 q_2}{N_2} + \frac{x_3 q_3}{N_3}\right)\right\} \qquad (4.3)$$

where $F_{q_1,q_2,q_3} = \left|F_{q_1,q_2,q_3}\right| * \exp\left(i * \varphi_{q_1,q_2,q_3}\right)$. The complex structure factor F_{q_1,q_2,q_3} has N unknown phases φ_{q_1,q_2,q_3} while $\left|F_{q_1,q_2,q_3}\right|$ are available from the X-ray diffraction experiment. Let's take $r = x_1 N_2 N_3 + x_2 N_3 + x_3$ and $u = q_1 N_2 N_3 + q_2 N_3 + q_3$, and $a_{r,u} = \frac{1}{N}\exp\left\{-2\pi i * \left(\frac{x_1 q_1}{N_1} + \frac{x_2 q_2}{N_2} + \frac{x_3 q_3}{N_3}\right)\right\}$.

Then we may also represent $N_1 \times N_2 \times N_3$ 3-D array ρ_{x_1,x_2,x_3} and F_{q_1,q_2,q_3} with size $1\times N$ 1-D array $\overrightarrow{\rho_r}$ and $\overrightarrow{F_u}$. Eqn (3.1) hence turns into linear equations

$$\begin{cases} \rho_0 = F_0 * a_{0,0} + F_1 * a_{0,1} + \cdots + F_{N-1} * a_{0,N-1} \\ \rho_1 = F_1 * a_{1,0} + F_1 * a_{1,1} + \cdots + F_{N-1} * a_{1,N-1} \\ \cdots \\ \rho_{N-1} = F_{N-1} * a_{N-1,0} + F_{N-1} * a_{N-1,1} + \cdots + F_{N-1} * a_{N-1,N-1} \end{cases} \quad (4.4)$$

We should note that complex coefficient matrix $\boldsymbol{a}$ is a known constant, which only depends on the number of sample values we took in real and reciprocal space, namely N_1, N_2, N_3. $\boldsymbol{a}$ is given by

$$\boldsymbol{a} = \begin{bmatrix} a_{0,0} & \cdots & a_{0,N-1} \\ \vdots & \ddots & \vdots \\ a_{N-1,0} & \cdots & a_{N-1,N-1} \end{bmatrix} \quad (4.5)$$

So the discrete Fourier transform can be reformulated into a set of linear equations.

$$\overrightarrow{\rho_r} = a_{r,u} * \overrightarrow{F_u} \quad (4.6)$$

As $\overrightarrow{F_u}$ and $\overrightarrow{\rho_r}$ are typically not completely known, so we may write (3.6) as

$$a_{r,u} * \overrightarrow{F_u} - \overrightarrow{\rho_r} = 0$$

In matrix form

$$(a_{r,u} \quad -I) * \begin{pmatrix} \overrightarrow{F_u} \\ \overrightarrow{\rho_r} \end{pmatrix} = 0$$

where I is an $N \times N$ matrix. If we have constraints that can be expressed as a set of linear equations in $\overrightarrow{F_u}$ and $\overrightarrow{\rho_r}$, then the coefficient matrix can be expressed as

$$\begin{pmatrix} a_{r,u} & -I \\ c_e & \end{pmatrix} * \begin{pmatrix} \overrightarrow{F_u} \\ \overrightarrow{\rho_r} \end{pmatrix} = \begin{pmatrix} 0 \\ c_r \end{pmatrix}$$

In this case, the uniqueness problem can be quantitatively analyzed by comparing the rank of the coefficient matrix and the augmented matrix. The system has a unique solution when the rank of coefficient matrix is equal to the number of augmented matrix. In particular, if the number of variables equals to the rank of coefficient, then the solution is unique. Otherwise, there are infinite solutions. If the rank of coefficient matrix is smaller than the rank of augmented matrix, then inconsistent equations are present, resulting in no solution.

Given a complete set of measured structure factor amplitudes $|\overrightarrow{F_u}|$, if no further information is available about the object density in real space or phases in reciprocal space, then

$\overrightarrow{\rho_r}$ needs to be treated as a complex number, with $2N$ unknown numbers in real part and imaginary part. Accordingly, we may write two equations for real part and imaginary part for each equation in (3.4), which gives us maximum $2N$ constraints if the rank of $\boldsymbol{a}$ is N. So the freedom of solution will be at least N. In this case, the solution is not unique.

If we have prior information that the object should be real, then we have N equations to constrain the imaginary part of $\overrightarrow{\rho_r}$ to be zero. We may write these constraints as

$$\begin{cases} imag(\rho_{0,0,0}) = 0 \\ imag(\rho_{0,0,1}) = 0 \\ \cdots \\ imag(\rho_{N-1,N-1,N-1}) = 0 \end{cases} \tag{4.7}$$

In addition, the real density also gives additional $N/2$ constraints on phases of structure factors by Friedel's law.

$$\begin{cases} \varphi_{0,0,0} = 0 \\ \varphi_{0,0,1} = 2\pi - \varphi_{0,0,\frac{N+1}{2}} \\ \cdots \\ \varphi_{\frac{N-1}{2},\frac{N-1}{2},\frac{N-1}{2}} = 2\pi - \varphi_{N-1,N-1,N-1} \end{cases} \tag{4.7}$$

It appears that we have $2N + N + \frac{N}{2} = 3\frac{N}{2}$ equations, which exceeds unknown variables $3N$. However, equations (4.6) and (4.7) are actually not independent to each other. Equation (3.6) is completely determined given equations (4.6). So knowing the object is "real" gives us actually $N/2$ independent constraints. Therefore, the freedom of solution is $2N - N - \frac{N}{2} = N/2$. The solution is still not uniquely determined. Nevertheless, the prior information reduced $N/2$ freedom, compared with no prior information. But still, the measured structure factor amplitude plus real object density doesn't give enough information about the real space. This can be shown in figure 4.1.

The following simulation shows that random phase that satisfies Friedel's law doesn't necessarily give correct model. Therefore, more constraints are needed to narrow down our searching possibility. So more knowledge is required to guarantee a unique solution.

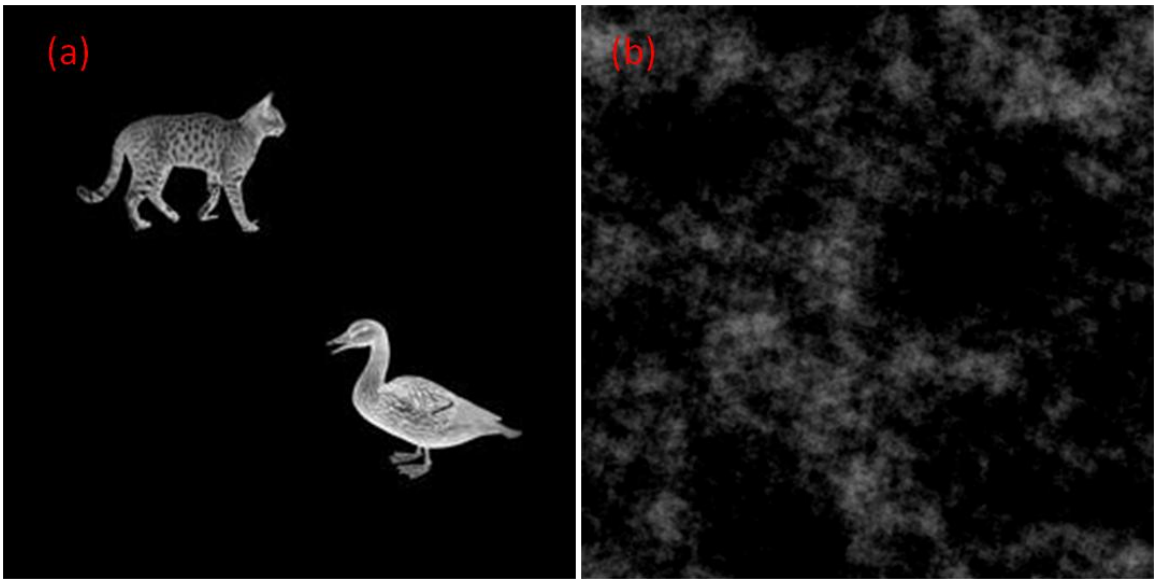


Figure 4.1 Random phase doesn't give correct structure. (a) is the model. (b) is the inverse Fourier transform of Fourier amplitudes with random phases which satisfy equation (4.7).

Additional constraints are required in real space or reciprocal phases to solve the structure. What if we further know there is a finite boundary of the real object, which is the case of non-periodic diffractive imaging experiments. If we know half of the real space information, then the solution is possibly unique. Then some delicate algorithm can find it. In crystallography, initial phases are typically obtained from model or inferred from experiment where protein is labeled with heavy atoms. In the rest of this chapter, we will demonstrate phasing a crystal diffraction dataset with various real space constraints using iterative projection algorithm.

4.2 Iterative projection algorithm

4.2.1 *Hybrid Input-Output algorithm*

HIO algorithm is developed from error reduction algorithm. In error reduction algorithm, the first object density is calculated by performing inverse Fourier transform on measured Fourier spectrum amplitudes and random phases. Then a real space constraint, called a support, is imposed on the density values which modifies values outside of support to zeros, while keep density values inside the support unchanged. Then new phases are estimated by performing

Fourier transform on the new object. To estimate the next object density, the measured amplitudes combined with phases estimated from the latest iteration are used in inverse Fourier transform. The flowchart is shown in Fig 4.2. The algorithm will converge to minima after certain number of iterations. One drawback of error reduction algorithm is that it is easy to be trapped in local minima. To solve this problem, the HIO algorithm is developed.

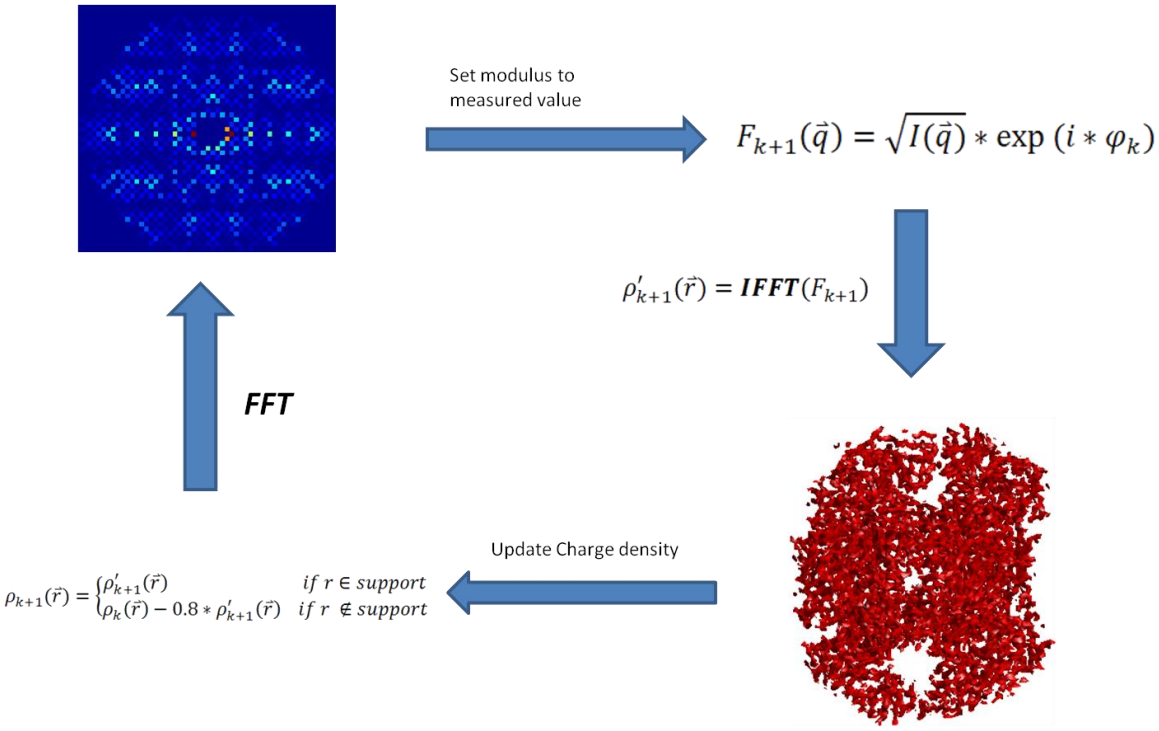


Fig 4.2 Hybrid Input-Output algorithm flowchart.

The implementation of the HIO algortihm is outlined in (Chapman et al., 2006; Spence, Weierstall, Fricke, Glaeser, & Downing, 2003). Here we briefly describe the procedure. HIO algorithm and error reduction is used to search the optimal solution. 10 error reduction steps are performed followed by 30 HIO-iterations, hoping to refine the structure. We assume that HIO algorithm can find global minima while error reduction can do further refinement. The number of iterations required for convergence depends on the molecular shape and envelope size.

To evaluate the iteration process, object space error metric in the kth iteration is introduced as (Spence 2003)

$$\varepsilon_k = \left(\frac{\sum_{(x,y,z)\in S} |\rho_k(x,y,z)|^2}{\sum_{(x,y,z)\in S} |\rho_k(x,y,z)|^2} \right)^{1/2}$$

where S is support, $\rho_k(x,y,z)$ is the electron density distribution at kth iteration.

As ε_k depends on type of proteins, we also introduce a relative error metric

$$\text{rms} = \varepsilon_k/\varepsilon_1$$

$$\text{rms} = \left(\frac{\sum_{(x,y,z)\in S} |\rho_k(x,y,z)|^2}{\sum_{(x,y,z)\in S} |\rho_k(x,y,z)|^2} \right)^{1/2}$$

where ε_1 is the image space error in first iteration.

The correlation coefficient between the true density and estimated density is equal to the normalized cross-correlation function at the origin (Spence 2003), given as

$$CC = \frac{\sum_h |F_h|^2 \cos(\phi_h^t - \phi_h^e)}{\sum_h |F_h|^2}$$

where F_h and ϕ_h^t are the true structure factor amplitude and phase, respectively. ϕ_h^e is the refined phase from iteration. We also introduced following error metric in our simulation.

Weighted phase error

$$< |\Delta\Phi| >_w = \frac{\sum_h (|F_h^t| + |F_h^e|) \arccos(\cos(\phi_h^t - \phi_h^e))}{\sum_h (|F_h^t| + |F_h^e|)}$$

Average phase error

$$< |\Delta\Phi| >_a = \frac{\sum_h \arccos(\cos(\phi_h^t - \phi_h^e))}{\sum_h 1}$$

Fourier Shell Correlation

$$FSC(k, k+\Delta k) = \frac{\text{Re} \sum_{[k,k+\Delta k]} F^t F^{e^*}}{\{\sum_{[k,k+\Delta k]} F^{t\,2} \sum_{[k,k+\Delta k]} F^{e\,2}\}^{1/2}}$$

R factor

$$R = \frac{\sum_h ||F_h{}^t| - |F_h{}^e||}{\sum_h |F_h{}^t|}$$

When a support in real space is given and finer sampling in reciprocal space is available, the number of equations will exceed the number of unknown variables. Each equation of the discrete Fourier transform can be considered as an elliptical surface in a higher dimension. The intersection of all these surfaces gives our possible solutions. The HIO algorithm starts from a random guess of phases. Then it approaches the solution by doing projections to the support in real space and amplitude constraints in reciprocal space iteratively (Marchesini 2007).

4.2.2 *Patterson function*

Patterson function is often used to solve the phase problem in crystallography. It is the inverse Fourier transform of intensities rather than structure factors

$$P(\vec{r}) = \sum_{\vec{u}} |F(\vec{u})|^2 e^{-2\pi i \vec{u}\cdot\vec{r}} \qquad (4.8)$$

Mathematically, Patterson function is equivalent to the autocorrelation of the object density, which is defined as

$$A(\vec{r}) = \sum_{\vec{r}'} \rho(\vec{r}') * \rho(\vec{r}' + \vec{r}) \qquad (4.9)$$

Here is a short proof. Inserting equation (4.1) to (4.9)

$$A(\vec{r}) = \int_{-\infty}^{+\infty} \left\{ \sum_{\vec{h}} F(\vec{h}) * e^{-2\pi i (\vec{h}\vec{r})} \right\} * \left\{ \sum_{\vec{k}} F(\vec{k}) * e^{-2\pi i \vec{k}(\vec{r}' + \vec{r})} \right\} d\vec{r}$$

$$= \sum_{\vec{h},\vec{k}} F(\vec{h}) F(\vec{k}) * e^{-2\pi i (\vec{k}\vec{r}')} \int_{-\infty}^{+\infty} e^{-2\pi i \vec{r}(\vec{h}-\vec{k})} d\vec{r}$$

$$= \sum_{\vec{h},\vec{k}} F(\vec{h}) F(\vec{k}) * e^{-2\pi i (\vec{k}\vec{r}')} \delta(\vec{h} - \vec{k})$$

$$= \sum_{\vec{h},\vec{k}} |F(\vec{h})|^2 e^{-2\pi i (\vec{k}\vec{r}')}$$

$$= P(\vec{r})$$

Patterson function is calculated from the Fourier spectrum while autocorrelation is calculated from real space object density. For single particle diffraction, the Patterson function is exactly the autocorrelation function, which is continuous to infinity. For X-ray crystallography, the

Patterson function is actually the autocorrelation of crystal, instead of individual unit cells [Ref to Rick Millane 2015]. The Patterson function is periodic with size L/2 in each dimension, which gives maximum half information in real space. Meanwhile, it contains more vectors in the L/2 region as it measures correlation between the different unit cells, as shown in figure 4.3.

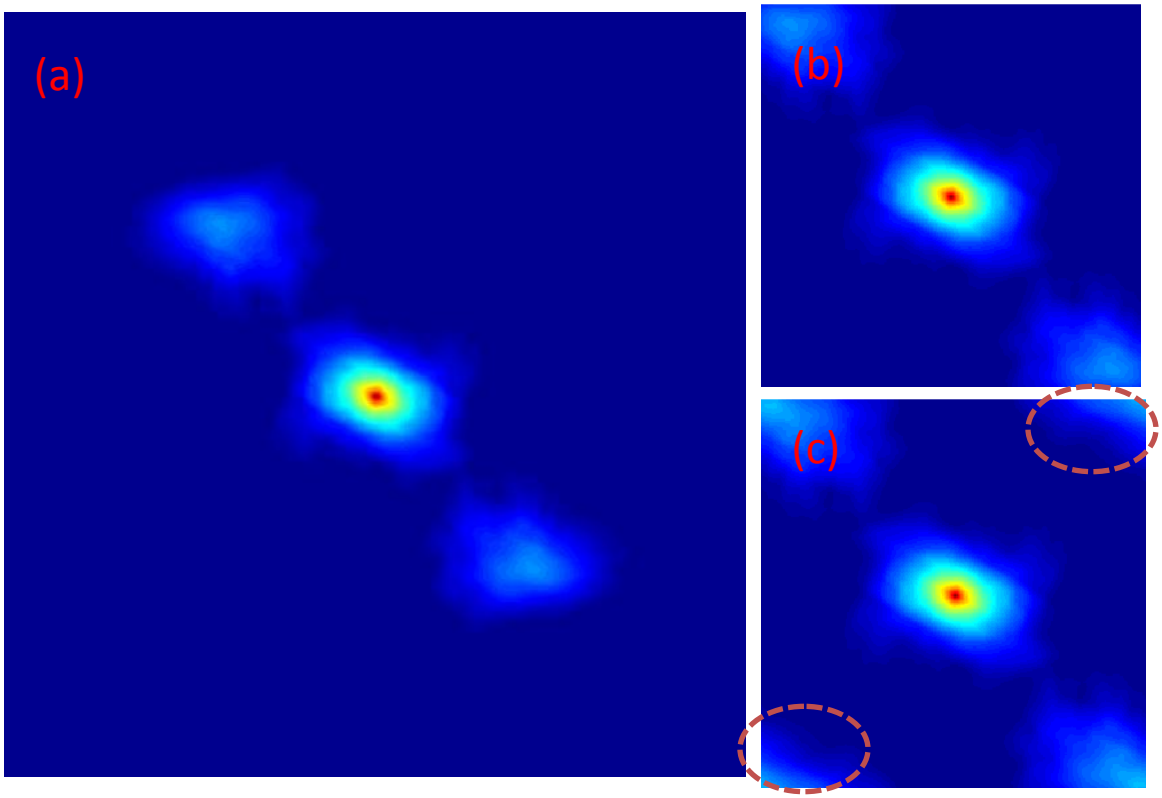


Figure 4.3 Autocorrelation function of isolated non-periodic object and its crystal form. (a) Autocorrelation of non-periodic object shown in Fig1. (b) Autocorrelation within L/2. (c) Patterson function of periodic object arranged in x and y diffraction.

4.2.3 *Resolution and oversampling ratio*

Resolution is one of the most concerned figures of merits in image processing. Its definition varies slightly across different imaging techniques, which is mainly due to the difference in experimental setup and data analysis. For example, in lens based optical systems, resolution is defined as the minimum separation of two points when the maximum intensity is 26% higher than the minimum between the two points. In X-ray diffractive imaging, the maximum resolution of a

diffraction pattern is given by the Bragg spots measured at maximum distance from center, which depends on X-ray wavelength and sample quality. The oversampling factor characterizes the minimum distance to obtain two discrete values in the reciprocal space, which has similar meaning as resolution in the real space. Quantitatively speaking, it often refers to the ratio between the inverse of object size and the minimum sampling space in reciprocal space. A larger oversampling factor means finer sampling in reciprocal space. For those diffraction patterns collected from scattered X-rays by non-periodic particle, the scattered intensity is continuous, where the oversampling factor is solely limited by the pixel size of detector.

Resolution in real space gives the maximum spatial frequency component in Fourier spectrum, while the oversampling ratio in reciprocal gives the maximum size of autocorrelation in real space. The oversampling ratio is often the key factor in phase retrieval for non-periodic objects because it gives information about the autocorrelation of the charge density function. Resolution gives the volume of the reciprocal space. In contrast, oversampling ratio gives the volume in real space. For the first statement, it is easy to understand that high resolution data means the presence of Bragg spots at high angle. We have a wider area of reciprocal space. A similar concept also applies to the oversampling ratio. If we sample finer, then we get bigger volume information of real space. This can be better illustrated in a numerical way.

Here we adopt a very straightforward definition of resolution. The minimum distance we can distinguish is our resolution limit. If we represent an object in a 2-D image, the minimum resolution is given by the pixel distance.

Given a protein molecule, its 3-D charge density map $\rho(\vec{r})$ is represented as a 3-D array $\rho[x]$, where $x = [x1, x2, x3]$. The size of the matrix is given by

$$N_{x1} = 2 * \frac{L_{x1}}{Res} + 1;\ N_{x2} = 2 * \frac{L_{x2}}{Res} + 1; N_{x3} = 2 * \frac{L_{x3}}{Res} + 1 \qquad (4.10)$$

where N_{x1}, N_{x2}, N_{x3} are the number of matrix elements in $x1, x2, x3$ dimension respectively; L_{x1}, L_{x2}, L_{x3} are the length of object in $x1, x2, x3$ dimension respectively; Res is the resolution of object. The value 'one' (unit constant) is added to make the Fourier space have central symmetry in equation (4.10). The Fourier transform of this matrix produces another 3-D

array with same size in Fourier space, which represents the complex structure factor amplitudes $F[k]$. Here structure factor is used as a generalized term for both periodic and non-periodic objects, referring to the function of spatial frequencies in Fourier space.

If we know enough real space information, we don't need to measure reciprocal space completely. So it would be fine if we miss some structure factors at high angle. We can treat them as free parameters. Under sampling means we assume that values in between are zeros. When we digitize continuous space, there should be infinite points. But when under-sampled, only a portion of these infinite points are considered, the rest aren't (and are treated as zeros).

4.2.4 *Supports*

Support is the indispensible part for an iterative phasing algorithm in coherent diffractive imaging. It provides the boundary constraints of objects in real space. Support values inside of the boundary are ones, while all the values outside are zeros. A correct support contains the entire object inside the boundary. A support is called tight support if it specifies the exact boundary of the object. Typically, the support is larger than the size of the object. More zero values in the support, more powerful the support is. In the extreme case when all values in the support are ones, no constraints are implemented by this support.

In a single particle diffraction experiment, there are mainly three ways to obtain a support: 1) its size information; 2) auto correlation function; 3) locate set. For an isolated particle, its size information is adequate to build a rectangular box support which contains the object. When the size information is not available, its autocorrelation function, which is the inverse Fourier transform of the square of structure factor amplitudes, also gives the boundary information of the object. To make an iterative phasing algorithm more efficient, it is often desirable to obtain a tighter support than the autocorrelation function support. More importantly, it is more likely to get a right solution with tighter support. In 1982 (J. R. Fienup, Crimmins, & Holsztynski, 1982), Fienup proposed a locate set theory to improve the autocorrelation support. The main procedure is as follows: 1) Find the extreme points (typically furthest) in a certain direction from the object support; 2) Move the support from autocorrelation to those extreme points, then we will get several

autocorrelation supports A1,A2,...; 3) Intersect A1,A2,A3 etc. The overlap region will be the compact support (as shown in figure 4.4).

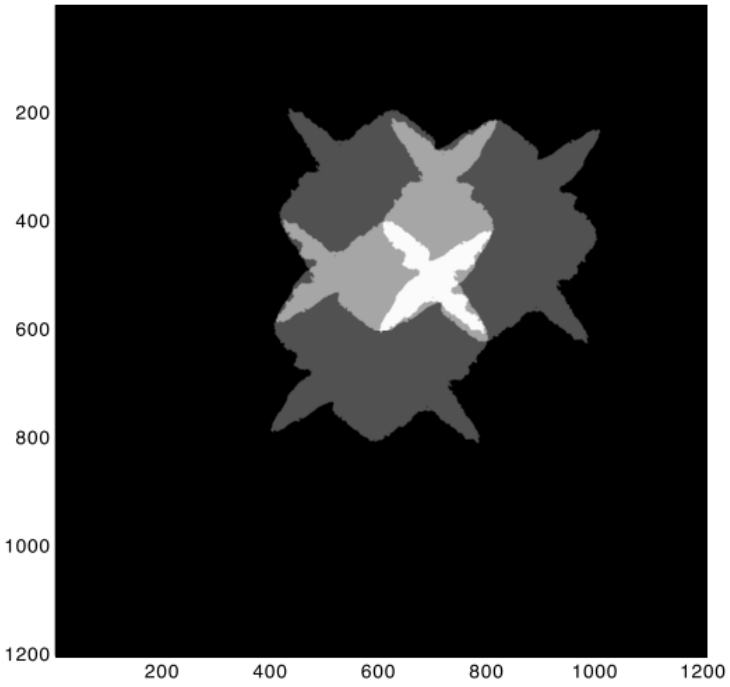


Figure 4.4 Locater set (James R Fienup, 2004)

In X-ray crystallography, much tighter support is required, compared with support in single particle imaging. First of all, the size information of unit cell doesn't provide additional constraints in real space since charge density outside the unit cell can't be set to zeros which is the case in single particle imaging. Secondly, the Patterson function calculated from crystal diffraction patterns is the autocorrelation function of the entire crystal, instead of an isolated molecule. It is periodic over the entire real space. In single particle diffraction, the Patterson function calculated from diffraction pattern is the autocorrelation of the object and its value is nonzero at a finite space. Moreover, inter vectors in crystal Patterson function reduces the room for imposing constraints in the autocorrelation support. It's also hard to apply locate set theory as an extreme set is difficult to obtain without any prior information. The extreme set consists several points on the object support boundary. But the shape of theprotein molecules is irregular. Actually,

the concept of "oversampling/under-sampling" doesn't fit to the realm of crystallography. An implicit assumption in oversampling is that the object should be non-periodic and finite in size. However, perfect crystals are considered to be periodic and infinite.

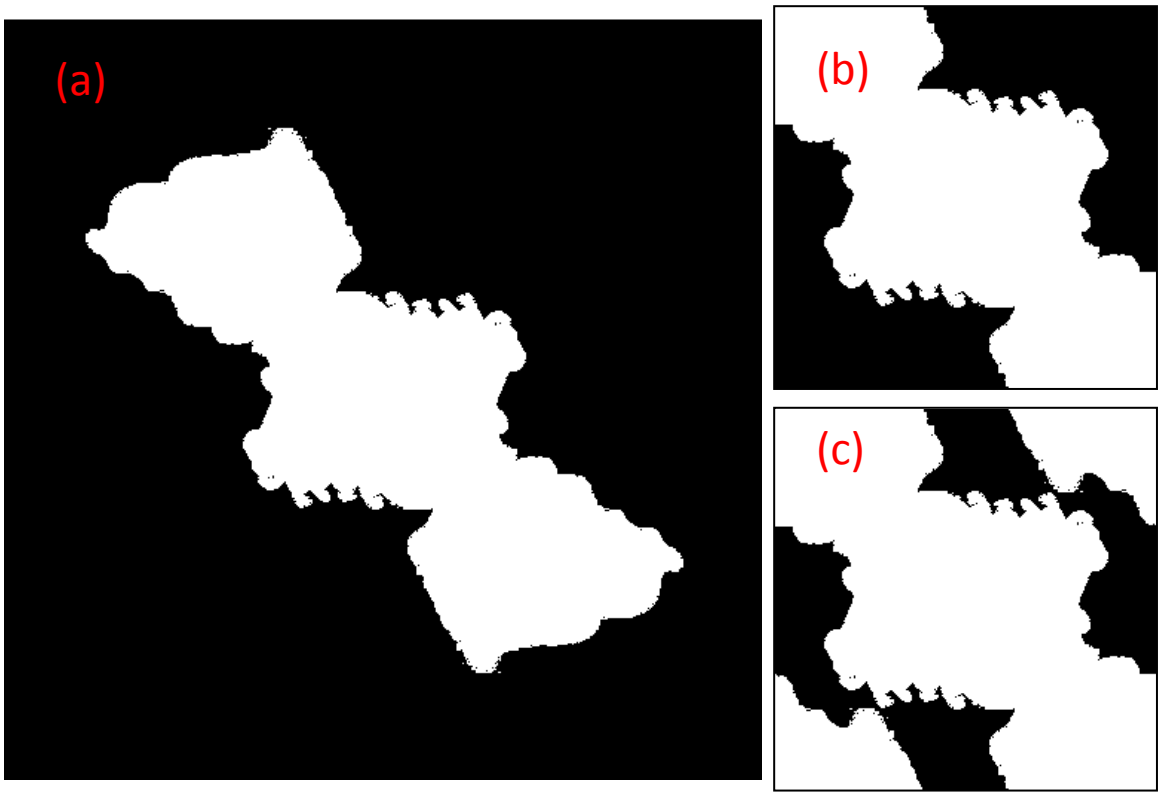


Fig 4.5 It shows support estimated from (a), Patterson function of isolated object, (b)half autocorrelation function in (c) Repeating unit of Patterson function of periodic object.

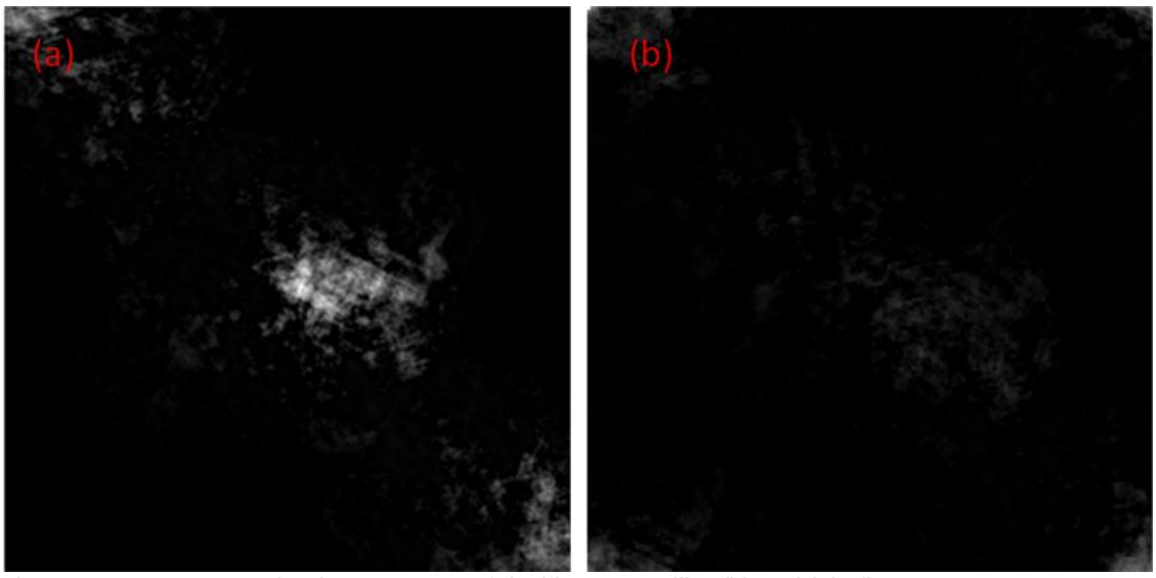


Figure 4.6 Reconstruction is not successful with support like (b) or (c) in figure 4.5

To facilitate iterative phasing algorithm in crystallography data, the shape information of the object is required i.e., the amount of vacuum inside of the unit cell need to be identified before applying algorithm.

In the following of this section, I will demonstrate several approaches to obtain a support and the image reconstruction with that support.

4.2.4.1 *Support from known object size*

In many X-ray single particle diffraction experiments, size information of the sample is often available. Even rough estimate of the object size is good enough to make a tight support for structure reconstruction. In the following example, the object is an image with cat and duck, which can be contained in 128*128 pixel box. Its diffraction pattern is oversampled by a factor of 2. The support can be created by designing a 2-D array with a square box with size 128*128 pixels in the center. All the values inside of the box are set to 1, and all the values outside the box are zeros. Its structure can be successfully reconstructed using HIO algorithm with support in Figure 4.9(b).

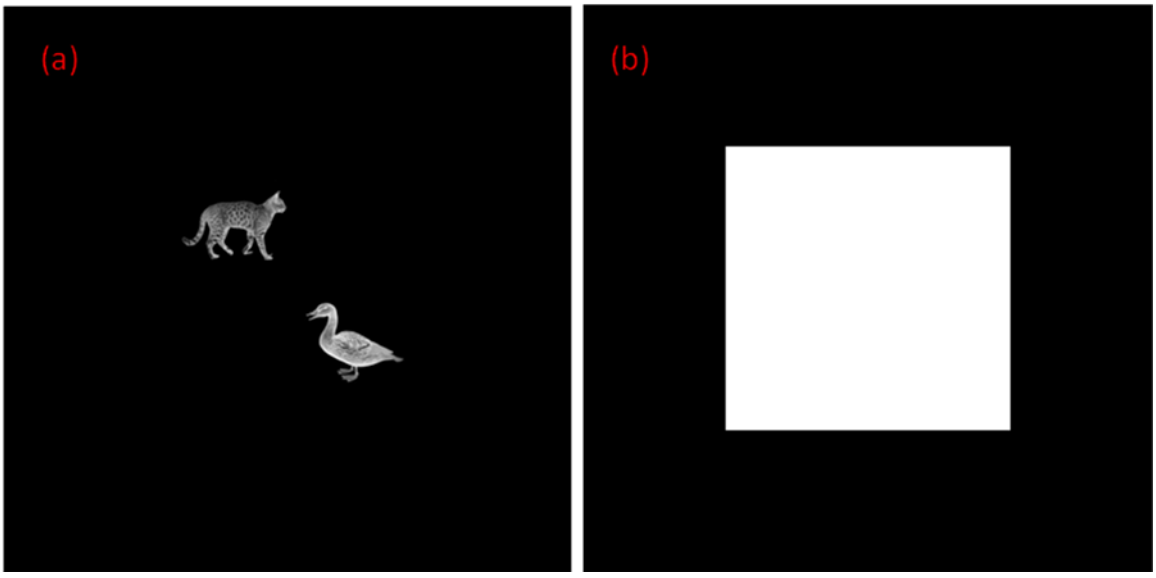


Figure 4.7 (a) object padded with zeros to 2X. (b) support from size information.

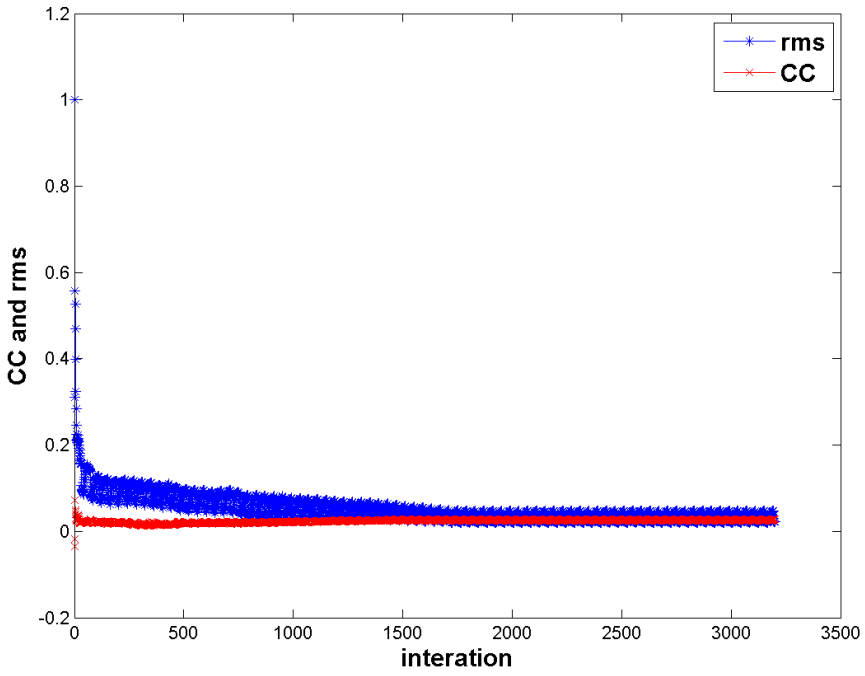


Figure 4.8 Correlation coefficient CC and rms value over iteration

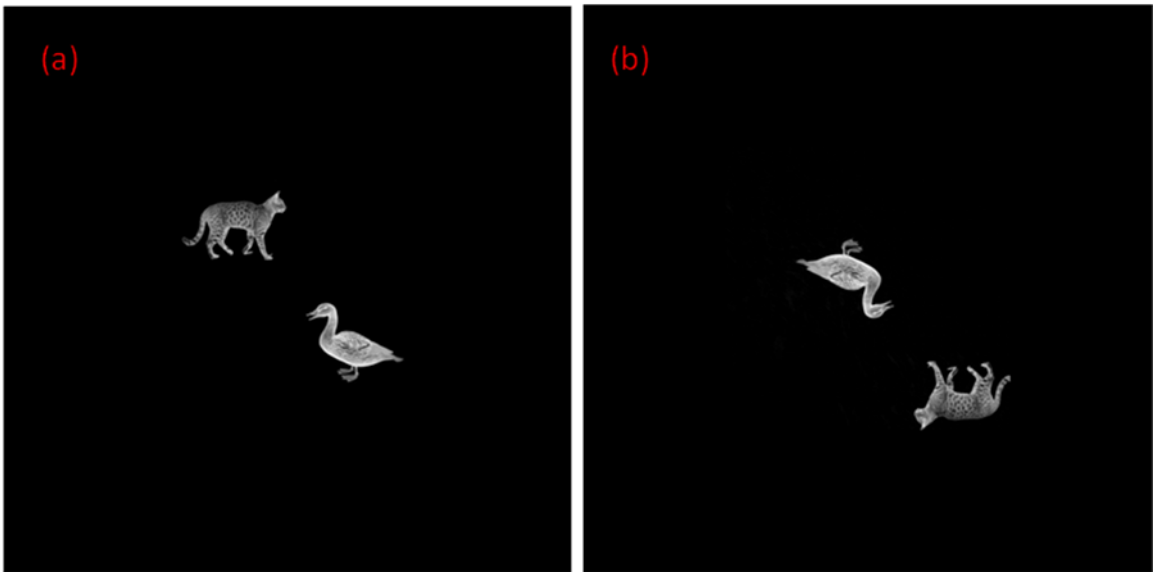


Figure 4.9 HIO reconstruction. (a) model (b) reconstruction with a inversion + translational shift. It happens in proteins as well. When size constraint is imposed, the reconstructions always seem inverted.

4.2.4.2 *Supports from autocorrelation function*

Object size information is not required for phasing single particle diffraction data. As long as the sample is finite, it is possible to derive a support and achieve structure reconstruction using autocorrelation functions. Autocorrelation shows all the vectors of intra-atom pairs within an object. So it spans maximum twice bigger than the original object in each dimension. The object should be contained by the outer boundary of autocorrelation function. Since all the translation and inversion of an object will give the same autocorrelation function, the support from autocorrelation function fits all such translations and inversions of the object. In other words, the right solution is not unique, but equivalent.

In the following simulation, the same object is used as in the previous example. The autocorrelation function $A(\vec{r})$ (shown in figure) is the same as $P(\vec{r})$ which is calculated by taking Fourier transform of the square of the structure factor amplitudes $|F(\vec{u})|^2$. Support is estimated from the autocorrelation function in the following way.

$$S(r) = \begin{cases} 1 & \text{if } A(r) > c \\ 0 & \text{if } A(r) \leq c \end{cases}$$

Here cutoff value c is constant. The value of c is zero in our simulation since no noise is introduced. The diffraction pattern is also oversampled by a factor of 2. The reconstruction with this support is shown in figure.

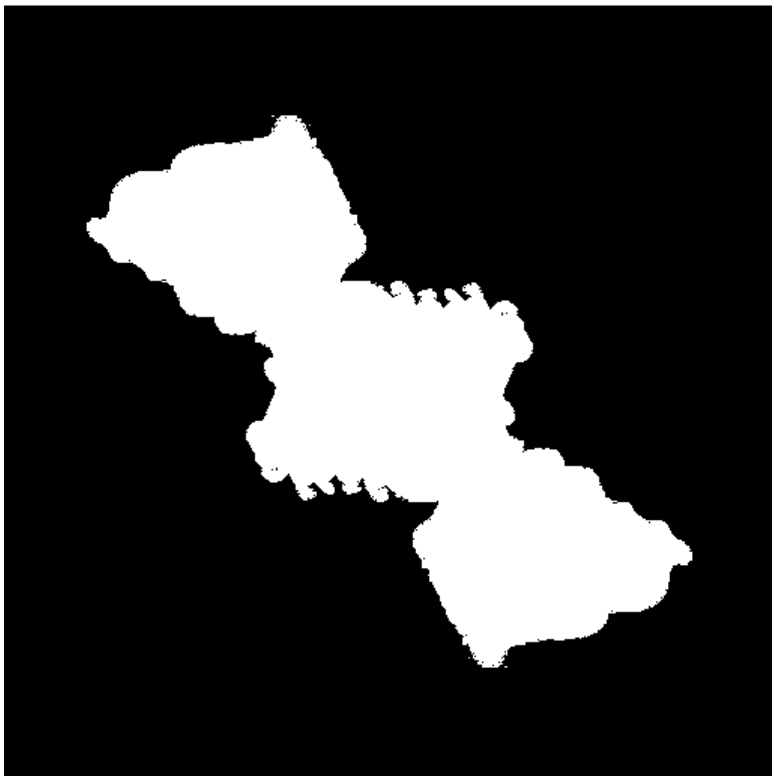


Fig 4.10 Support from Autocorrelation function

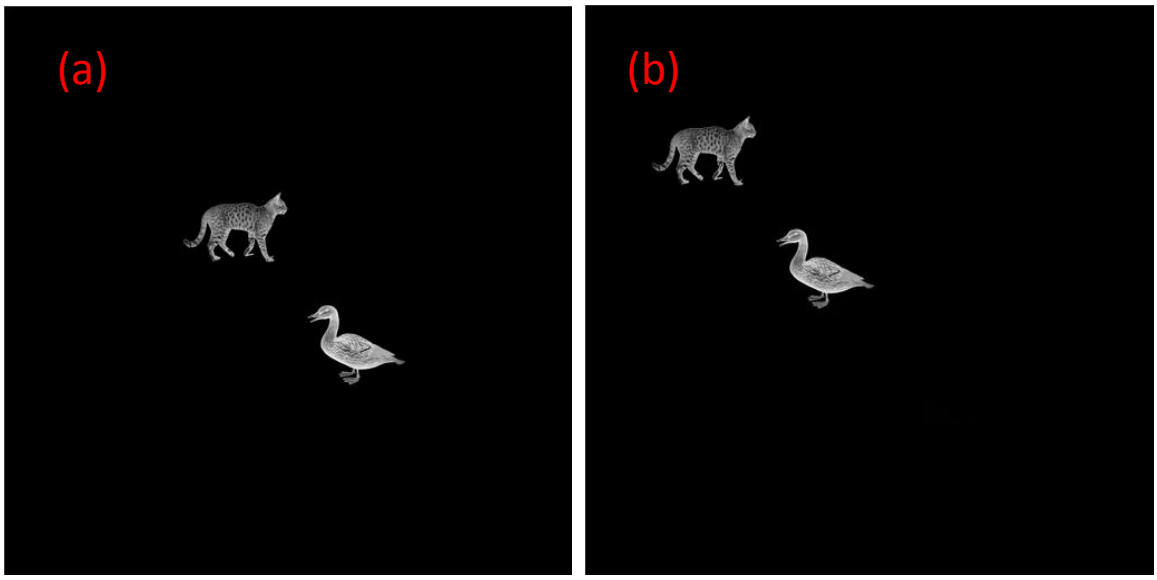


Figure 4.11 Model (a) and reconstruction (b) with a origin shift.

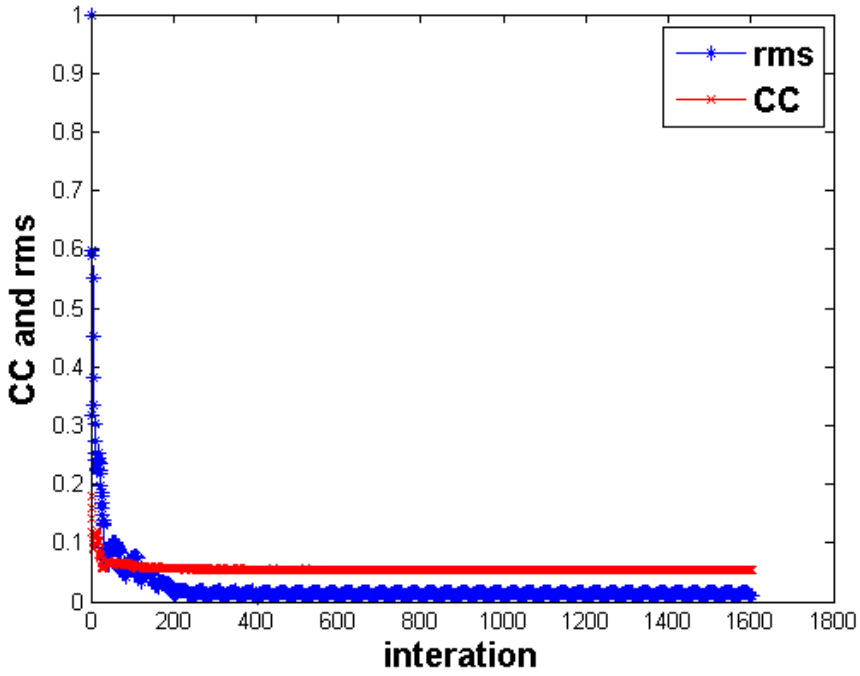


Figure 4.12 Rms over iteration. CC is low because of the origin shift.

4.2.4.2 *Support for periodic object*

Finite size constraints doesn't apply for periodic object, such as crystals. The charge density outside of unit cell can't be assumed to be zero as its adjacent are unit cells with same charge density distribution. The support from crystal density autocorrelation imposed very few constraints because the existence of inter vector between unit cells. Therefore, a much strong support are needed for phase retrieval. In the following simulation, the unit cell contains a cat and duck (Fig 4.13a). The rest black region are all zeros. Suppose we have a rough estimate of the boundary of cat and duck, then this support (Fig 4.13b) can be used to retrieve phases and reconstruct its origin image (Fig 4.14) with structure factors, using HIO algorithm described in Fig 4.2.

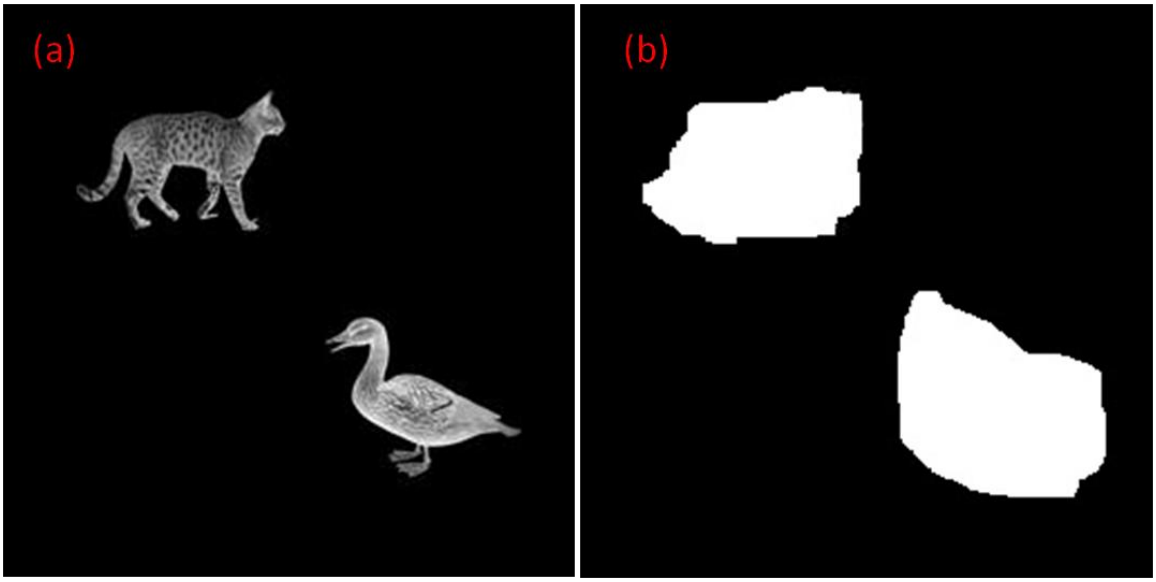


Figure 4.13 unit cell and internal support.

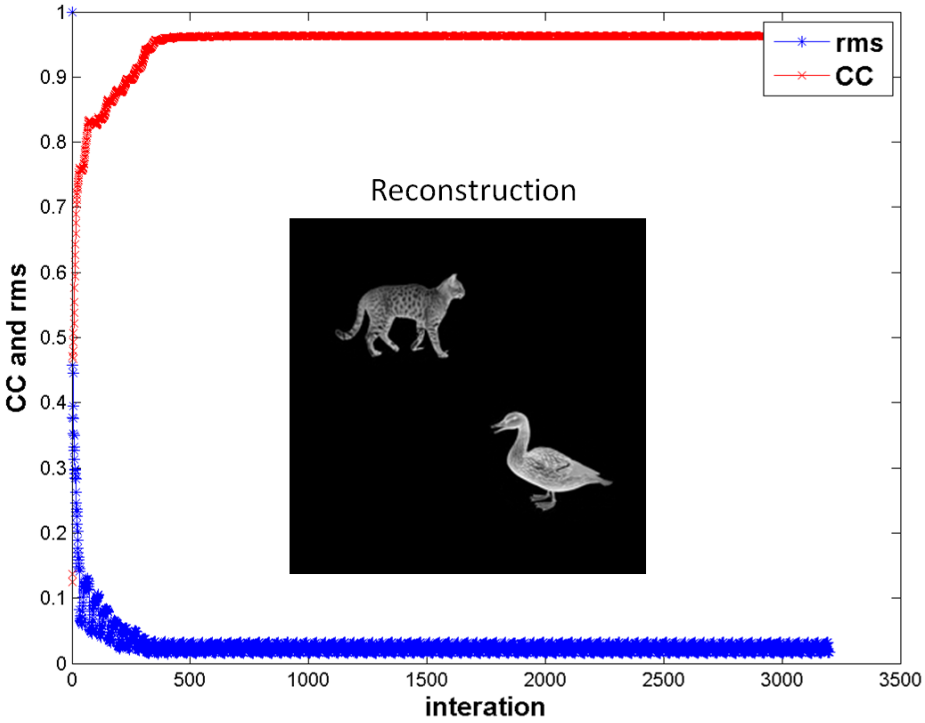


Figure 4.14 Reconstruction from tight support.

4.3 Application to two-dimensional streptavidin crystal diffraction data

4.3.1 *Streptavidin*

To illustrate the phasing algorithm for a realistic example, we choose two-dimensional streptavidin crystal as our model system. The first X-ray diffraction dataset from two dimensional streptavidin crystals was collected by Matthias et al using femtosecond X-ray pulses from an X-ray free electron laser (XFEL) (Frank et al., 2014). It was not possible to acquire transmission X-ray diffraction pattern from individual 2-D protein crystals at synchrotron due to radiation damage.

Streptavidin is a 52.8 kDa protein purified from the bacterium *Streptomyces Avidinii*. Streptavidin homotetramers have an extrodinary high affinity for biotin. With a dissociation constant on the order of 10^-14 mol/L, the binding of biotin to streptavidin is one of the strongest non-covalent interactions known in nature. Streptavidin is used extensively in molecular biology and bionanotechnology due to the streptavidin-biotin complex's extremes of temperature and pH (wiki).

The protein streptavidin is one of the most widely used proteins in molecular biology, biotechnology, and more recently, nanotechnology. The interaction between streptavidin and its natural ligand, biotin, is one of the strongest non-covalent interaction in biology (Kd ~10^-14) (Magalhães et al., 2011). As a result, this protein ligand couple has been the subject of numerous investigations to understand the nature of high affinity protein interaction as well as the target of multiple engineering efforts to alter its specificity and/or binding properties.

Charge density for two-dimensional crystal is periodic along lateral direction while continuous in its normal direction. The charges above and below the monolayer can be considered to be zeros. Hence, the reciprocal space is composed by a set of rods. In contrast to 3D crystals, the intensity is continuous in the normal direction.

4.3.2 *Phasing with compact support alone*

We first applied this algorithm on streptavidin (pdbid: 3RDX). The unit cell is orthorhombic, C 2 2 21, with cell constants $a = 79.25$ Å, $b = 81.64$ Å, $c = 84.54$ Å. We generate all atoms to fill the unit cell by symmetry operation as defined in the pdb file. Then the new pdb file was used in sFALL to calculate the structure factors of the unit cell to 3Å resolution in P1 symmetry. Subsequently, the structure factors are expanded to full reciprocal space by the following relation

$$I(h,k,l) = I(h,k,-l)$$

$$\varphi(h,k,l) = 2\pi - \varphi(h,k,-l)$$

The electron density map of one unit cell is generated by inverse Fourier transform of the full reciprocal space, which is a $59 \times 59 \times 59$ matrix in Matlab. Then we pad zeros above and below the unit cell along c axis to get a triple cell, with dimension $59 \times 59 \times 177$. The complex structure factors can be extracted by Fourier transform of the triple cell in Matlab.

The 3D support matrix is set to unity within the monolayer protein and zeros elsewhere. The structure factor amplitudes are used as a constraint in reciprocal space. The starting phases in HIO algorithm are random. The iteration in our algorithm consists of a 20 HIO sequence followed by a 20 error-reduction sequence.

Without any prior phase information, CC value converges very fast during the first 10 iteration steps. After around 20 iterations, the rms value decreases very slowly while CC value converges to 0.688, as shown in Fig 4.15. However, the promising rms and CC value doesn't give good estimate of 3D structure, as shown in fig 4.16. The side view along a and b axis resembles model slightly, but the density is far off along c axis projection.

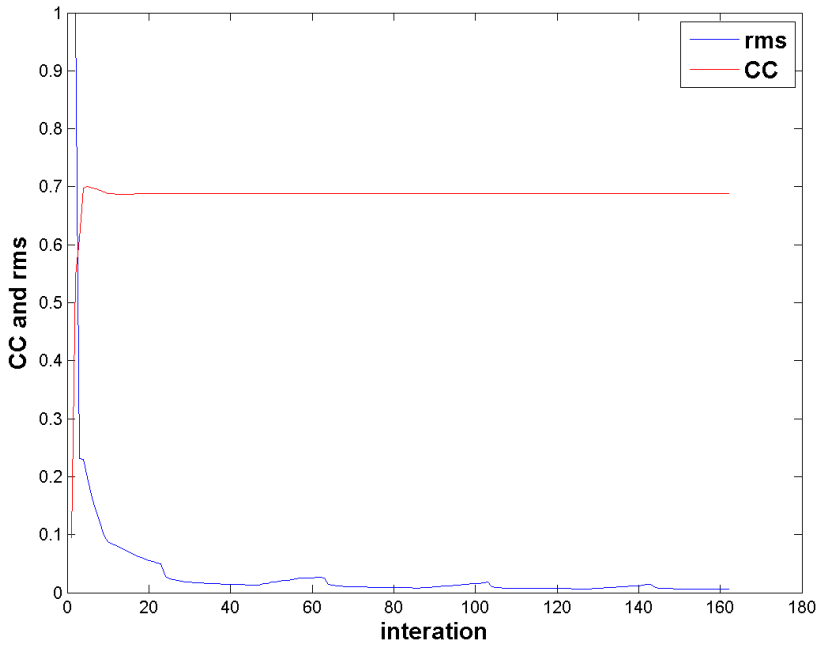


Fig 4.15 Correlation coefficient CC and rms value over iteration, starting with known structure amplitudes and random phases

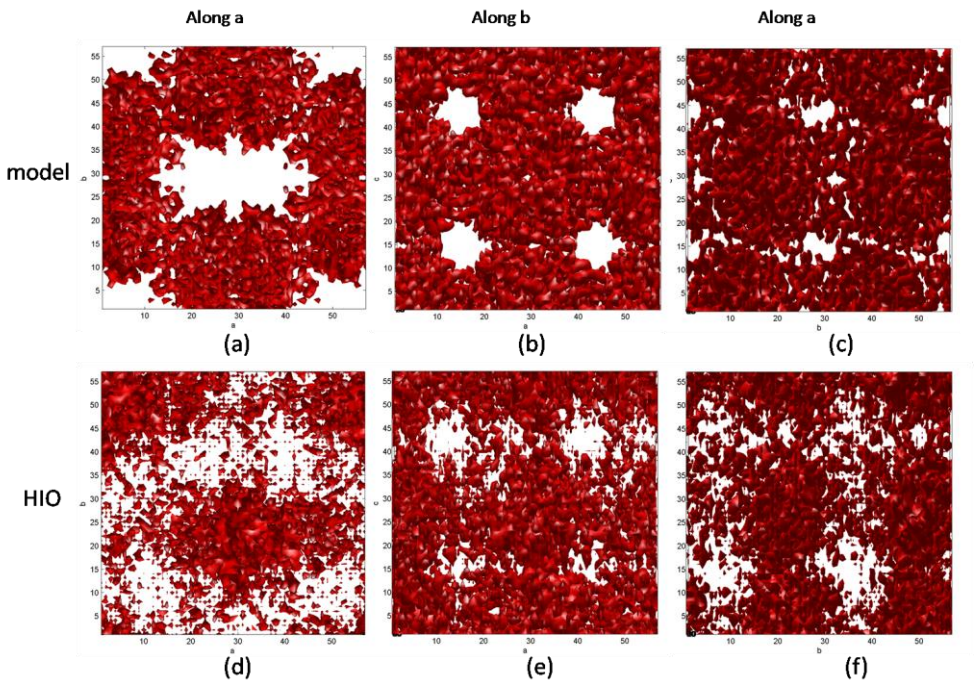


Fig 4.16 Comparison between model and structure from HIO estimate. (a-c) shows the model density along a, b, c axes; (d-f) shows the HIO estimated density view along c, b, a axis.

A careful examination of the projection along a and b axis direction shows that there seems to be a shift along certain directions. In order to verify whether the structure was recovered along c axis, we calculated the electron density projection from a-b plane on c axis, as shown in Fig 4.17. There is a high correlation between the model and the structure from HIO estimate. It seems that the density in real space is successfully reconstructed in c direction, which is equivalent to 1D phasing.

As we only oversample reciprocal space in c direction, the phases along each rod in reciprocal space were determined independently. But the relative phase between each rods in reciprocal space were not balanced during HIO algorithm. As a result, the inverse Fourier transform of each rod from reciprocal space gives the right rods along c direction in real space. But they are seated randomly in real space a-b plane. Phasing a 2-D crystal diffraction dataset is equivalent to 1D problem.

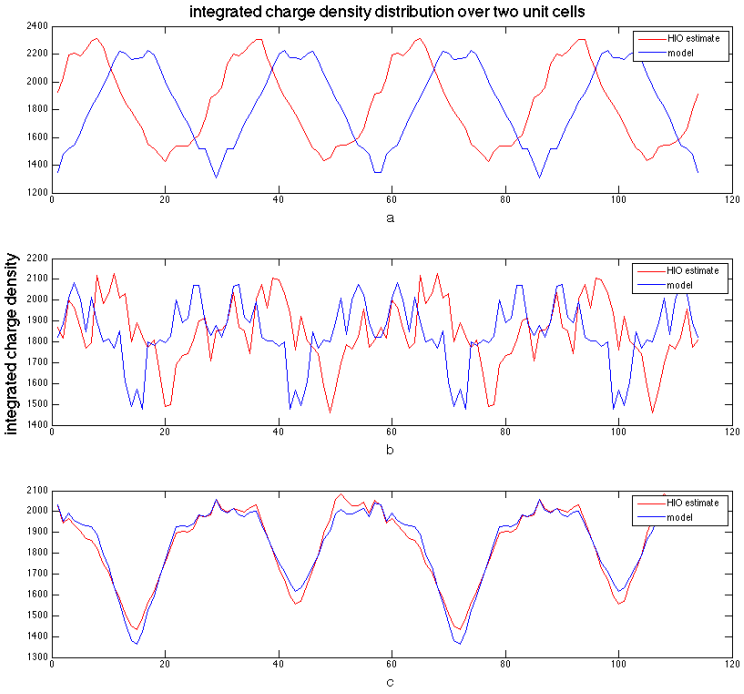


Figure 4.17 A comparison of density projection on c axis between model and structure from HIO estimate, without any prior phase information.

4.3.3 *Phasing with point support*

Protein crystals typically contain a large portion of disordered solvent. Although charge densities of the solvent are comparable with protein molecules, their contribution to the diffraction is much weaker than the signal from ordered protein molecules at Bragg spots. As we discussed in chapter 1, the signal at Bragg spots will be amplified by N^2 if N molecules are arranged in order along a specific dimension. Ideally, after background subtraction, the noise from instrumentation as well as solvent will be eliminated. Therefore, we may approximate the solvent region in unit cell as vacuum while simulating diffraction patterns. This approximation provides further constraints for HIO algorithm and opens up the possibility to phase the diffraction dataset with a known solvent-protein boundary.

Although solvent fraction is provided in PDB file, the boundary between solvent and protein is not specifically described. The charge density map generated from PDB file purely

shows contribution from protein molecules, not the real charge density of unit cell which contains solvent. However, it is safe to assume that the regions with low charge density value calculated from PDB are occupied by solvent molecules. Actually, protein molecule features are typically shown at a contour level between 1 to 3 sigma above average charge density, which accounts 5~20% volume of unit cell. If the display contour level is too low, it is very likely we only see a blob without any detailed features.

To study how much known solvent is required for successful reconstruction, we need to modify the density map $\rho(r)$ directly generated from PDB so that there is an explicit boundary between solvent and protein molecule. The most straightforward way would be set all charge density in original model $\rho(r)$ below a certain cut off ρ_0 to zero, which is considered as solvent region.

$$\rho'(r) = \begin{cases} \rho(r) & \text{if } \rho(r) \geq \rho_0 \\ 0 & \text{if } \rho(r) < \rho_0 \end{cases} \qquad (4.11)$$

Here $\rho'(r)$ is the charge density of new model, ρ_0 is the cut off density with units of sigma level. It may be set at certain value as long we may see satisfactory feature, such as an alpha helix, beta sheet or even benzene rings.

Accordingly, we may set a solvent support for our object with the similar idea.

$$S(r) = \begin{cases} 1 & \text{if } \rho(r) \geq \rho_s \\ 0 & \text{if } \rho(r) < \rho_s \end{cases} \qquad (4.12)$$

If electron density is lower than the cut off, then we set the support at this voxel as zero. Otherwise, the voxel value will be one. Those voxels with zeros values are prior information about solvent. More the number of zero-values, more the constraint imposed by this support. However, ρ_s needs to be smaller than ρ_0 to avoid conflict between support and model. Otherwise, the support will enforce some non-zero voxel values in protein region as defined in (3.11) to zero. The support is called tight support if $\rho_s = \rho_0$ and loose support if $\rho_s < \rho_0$. Here we call it point support in general, because the support gives precise solvent voxel points instead of a continuous and connected volume (as shown in figure).

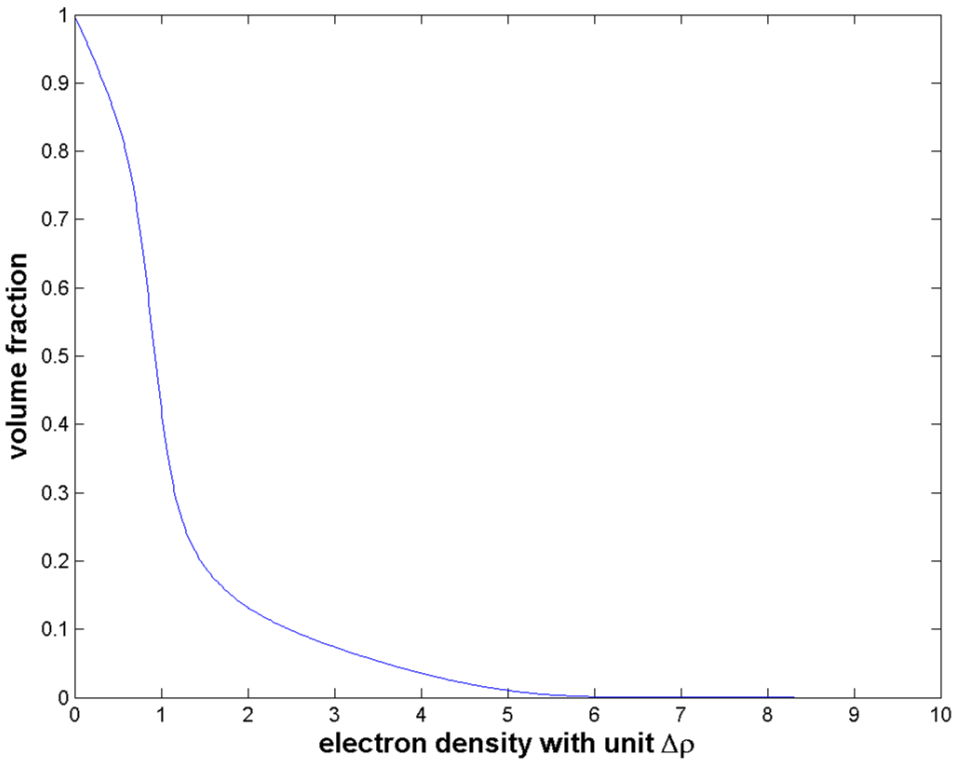
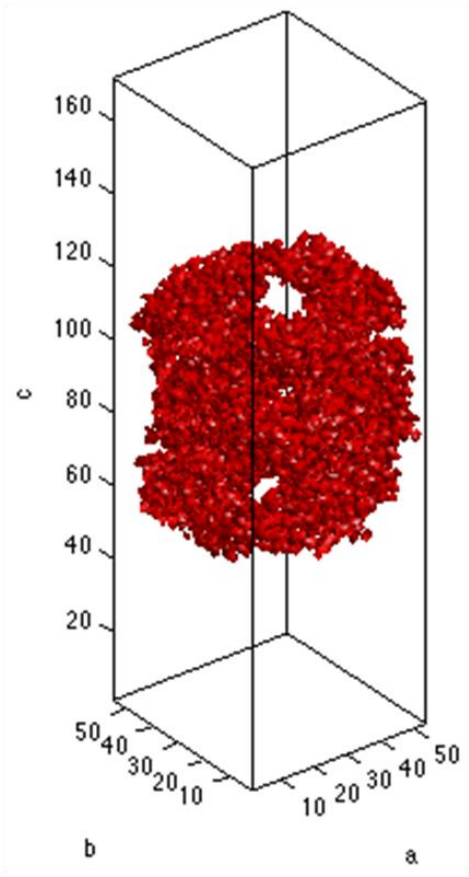


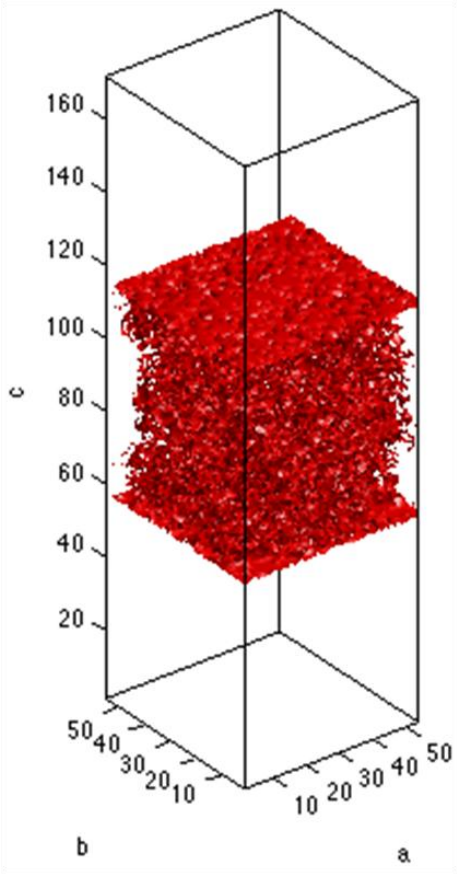
Figure 4.18 Volume fraction over cut off density value.

In the following simulation, we take $\rho_0 = 2 * \Delta\rho$, $C = 0$, while $\rho_s = 0.7 * \Delta\rho$. In this case, only 26.9% percent of unit cell volume is occupied by the object. Non-zero values in the support make 73.1% of a unit cell. Apart from this, we also apply a compact support above and below the unit cell. A decent structure reconstruction was obtained from the intensities, without the appeal for lateral oversampling in reciprocal space, shown in Fig 10,11,12.

We don't need a very tight support for this case. Our support identifies 17.9% voxel values of unit cell, which are known zeros. Such a support could potentially be obtained from a low-resolution image or Patterson function.

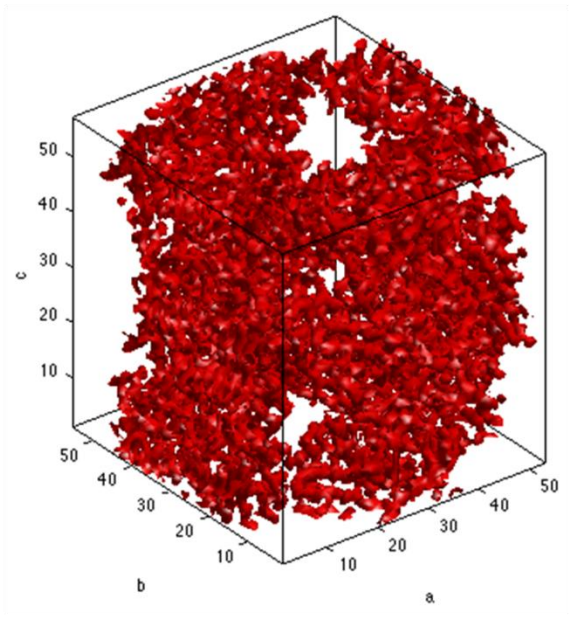


(a) Triple cell after solvent flatting

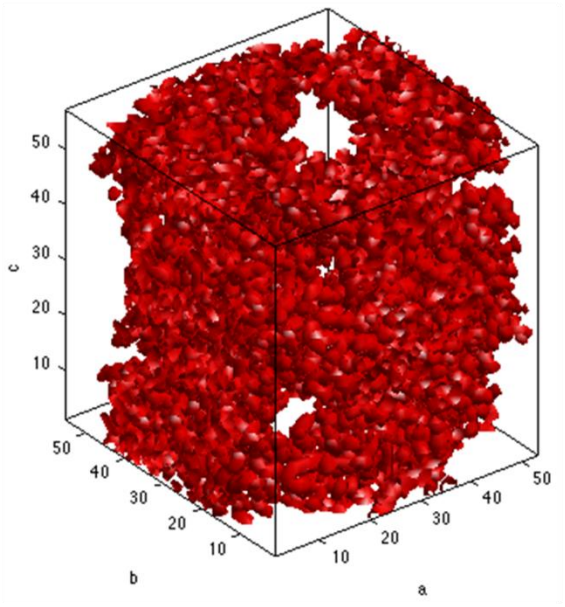


(b) Support

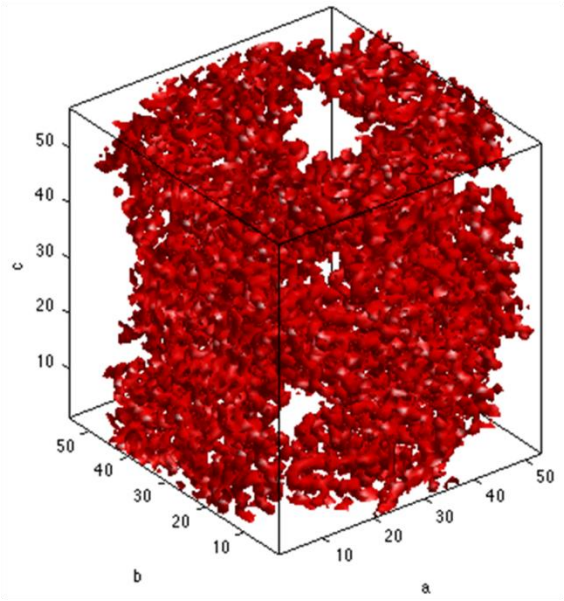
Figure 4.19 We apply solvent flatting both on model and support. Here ratio of non-zero to zero volume in support is about 3.



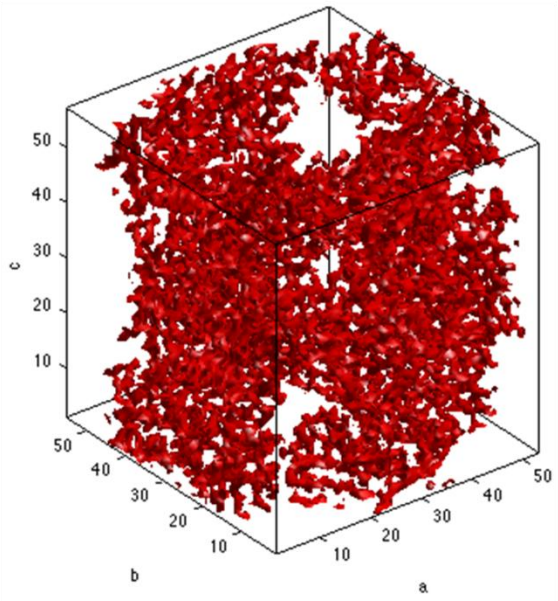
(a) Model cut at 2 sigma



(b) HIO estimate at 0.01 sigma



(c) HIO estimate at 1 sigma



(d) HIO estimate at 2 sigma

Figure 4.20 A comparison between model and HIO estimate shown at different sigma levels.

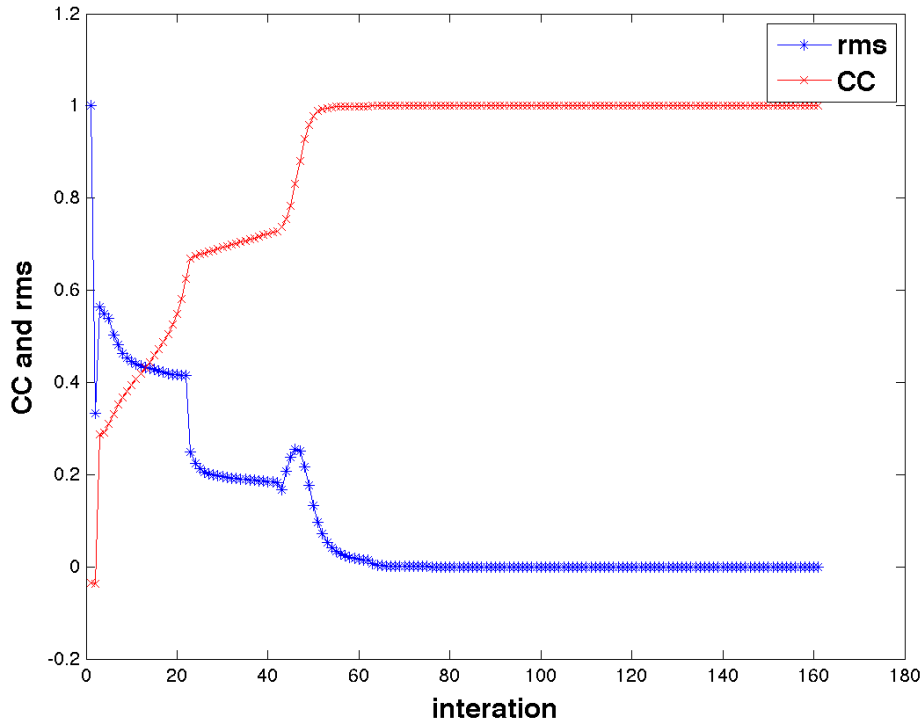


Figure 4.21 Correlation coefficient CC and rms value over iteration. It takes approximately 60 iterations to converge.

4.3.4 *Phasing with molecular envelope*

The point support is a very strong support since it identifies even small vacuum voxels or solvent location inside the protein pocket. Given the same volume of identified solvent, point support renders maximum independent constraints over other type of supports discussed in the following section. The simulation above provides the theoretical upper limit of solvent content required for unique phase retrieval. However, obtaining such a point support is highly dependent on atomic model, which is not practical for real data analysis. Here I demonstrate that phase retrieval is achievable given a roughly accurate molecular envelope. This envelope is connected and continuous volume in real space which contains protein molecules.

In the following simulation, a Gaussian filter is applied first first before defining the contour. Second, the Gaussian filter is applied to the point support. It will only enlarge the support area.

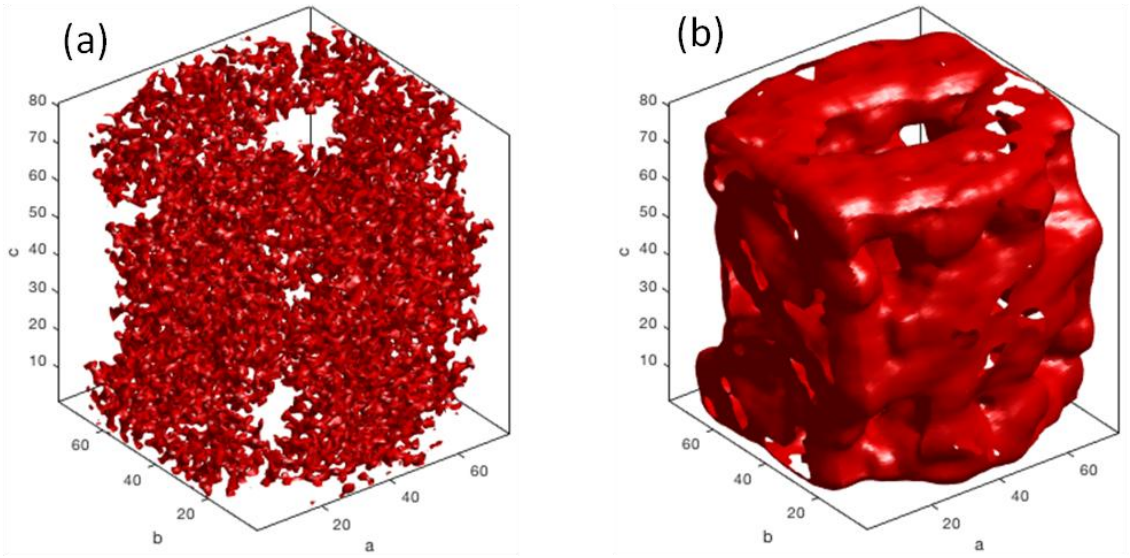


Figure 4.22 model and its rough molecular envelope

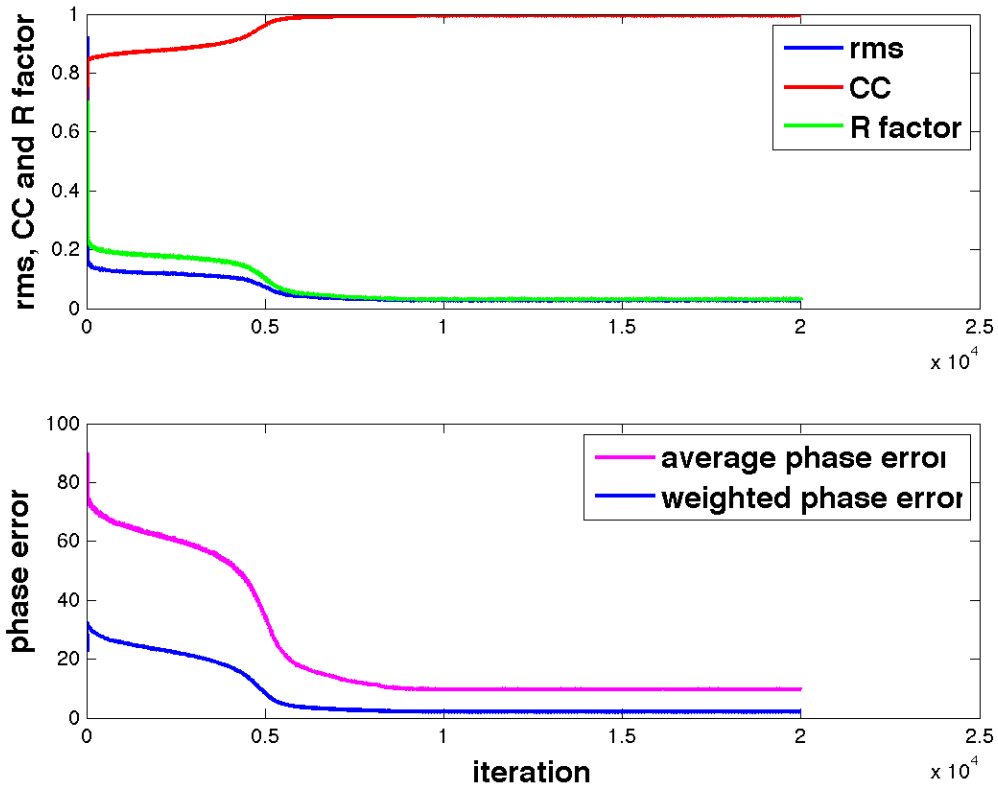


Figure 4.23 Correlation coefficient CC and rms value over iteration.

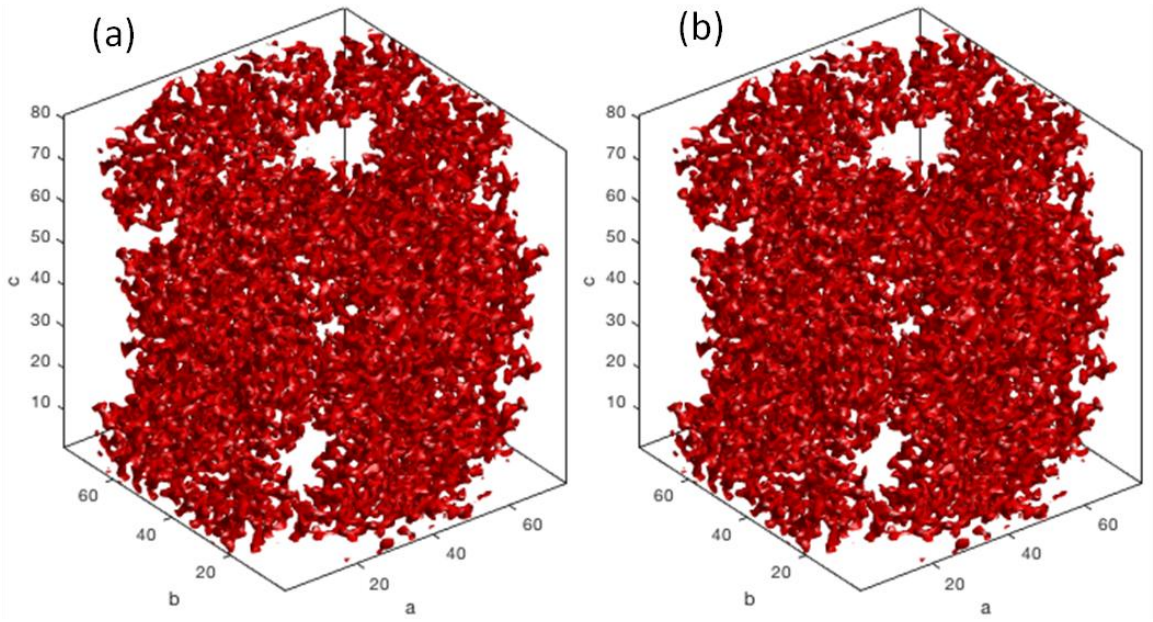


Figure 4.24 A comparison between original model (a) and reconstruction (b).

4.3.5 *Omit map implementation with IPA*

Omit map is widely used for reducing model bias. In the conventional procedure, a small region of model are systematically excluded for refinement. If there is no model bias, the density map calculated from experimental structure factor amplitudes with refined phases should reveal the missing region of model which is used for obtaining phases. Using HIO algorithm, a new approach is developed to validate the model.

In previous section, a point support can be estimated from atomic model. To validate the structure a small region of the molecule, a new support can be designed so that all constraints in those region are removed. If the model is correct, then the omit region should be fully recovered with experimental structure factors and the support estimated from structure with omitted region. In the following simulation, we create a support from a model with omit region which is a rectangular block as shown in figure 4.25(b). Using HIO algorithm, the structure of omitted region can be exactly recovered with this support and structure factor amplitudes(as shown figure 4.26)

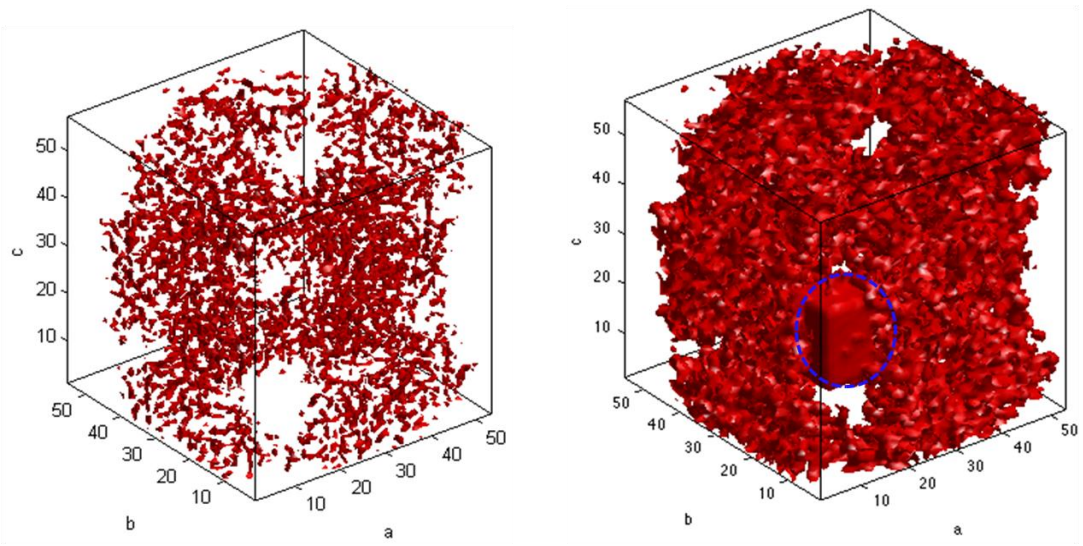


Figure 4.25 (a) model and (b) support estimated from model with omit region

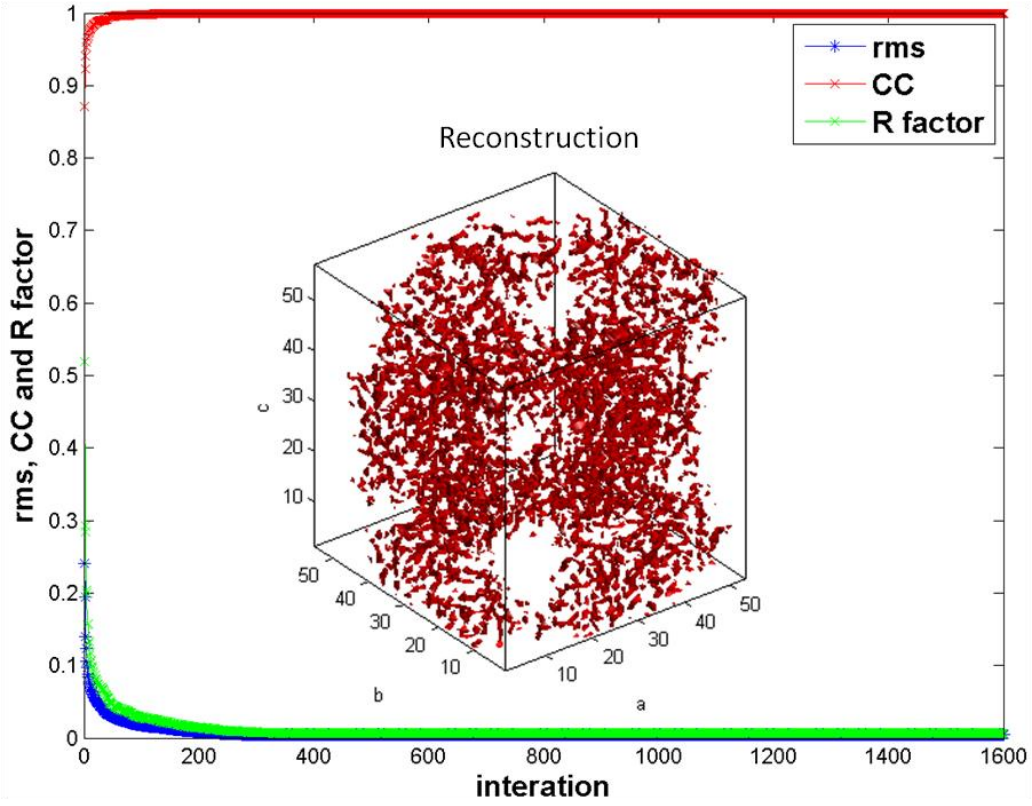


Figure 4.26 Correlation coefficient CC and rms value over iteration

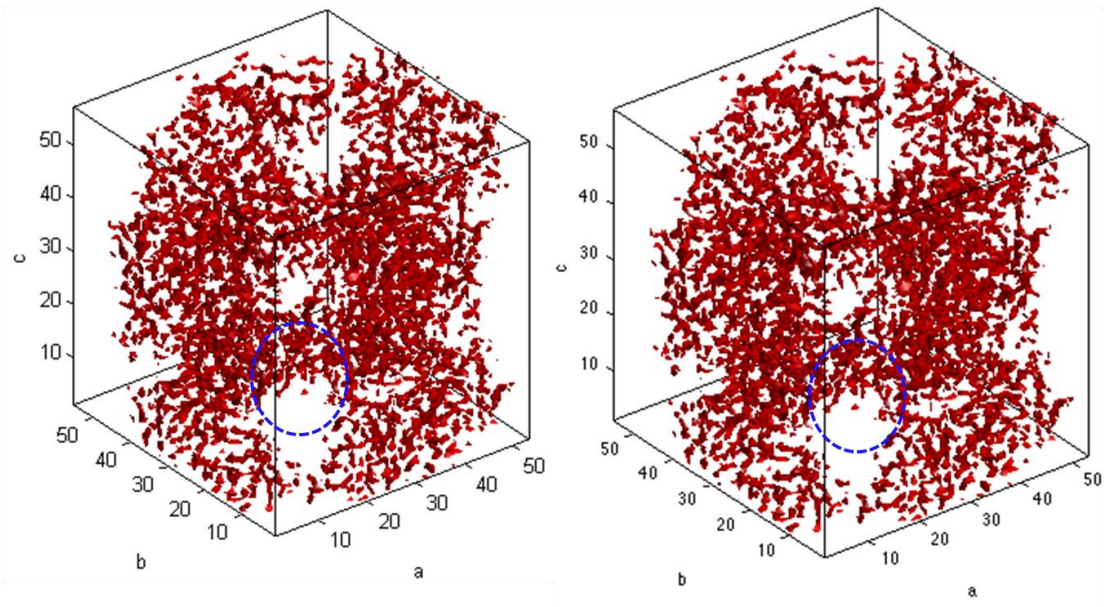


Figure 4.27 Comparison between model and HIO reconstruction

4.3.6 *Molecular replacement implementation with IPA*

In conventional molecular replacement, the phases are directly estimated from a model. The first electron density map is created using experimental structure factors and the phases calculated from model. Since phases carry more structure information, this method will introduce significant bias in our initial phase estimate. Therefore, this method is only used on the assumption that target structure is very similar to the model for phasing.

IPA facilitates an alternative way to do molecular replacement, which is more direct to the similar structure assumption. The similarity in molecular shape does not necessarily result in very similar phases. Phase error can be very high at high resolutions. In this new approach, only the shape information of model is used for create a support. If the support correct contains the target molecule, then it is possible to reconstruct structure free from error given perfect structure factor amplitudes.

In the following simulation, we choose streptavidin complexed with PEG (pdb:3rdu) as our target molecule. And 100% complete x-ray diffraction structure factor amplitude to 3A are simulated. A model of streptavidin free from ligand is also available (pdb:3rdx). To solved the

structure, a support is firstly estimated from model 3rdx (4.29b). Using IPA algorithm, the structure is solved in figure 4.31b, which more like our target model, instead of model for phasing.

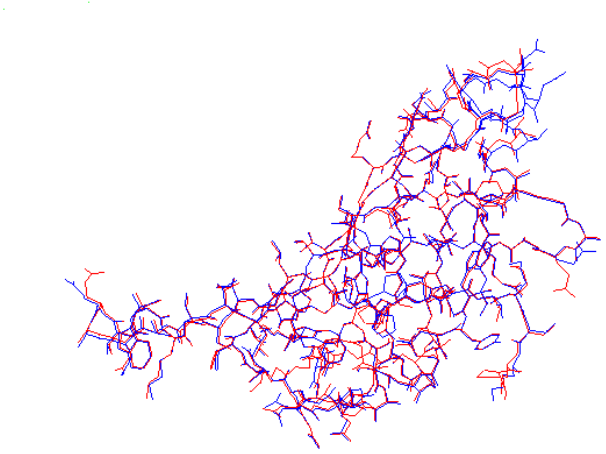


Figure 4.28 Comparison between two models. blue-3rdx,red-3rdu (streptavidin complexed with PEG). Only one monomer. We will phase 3rdu with 3rdx model.

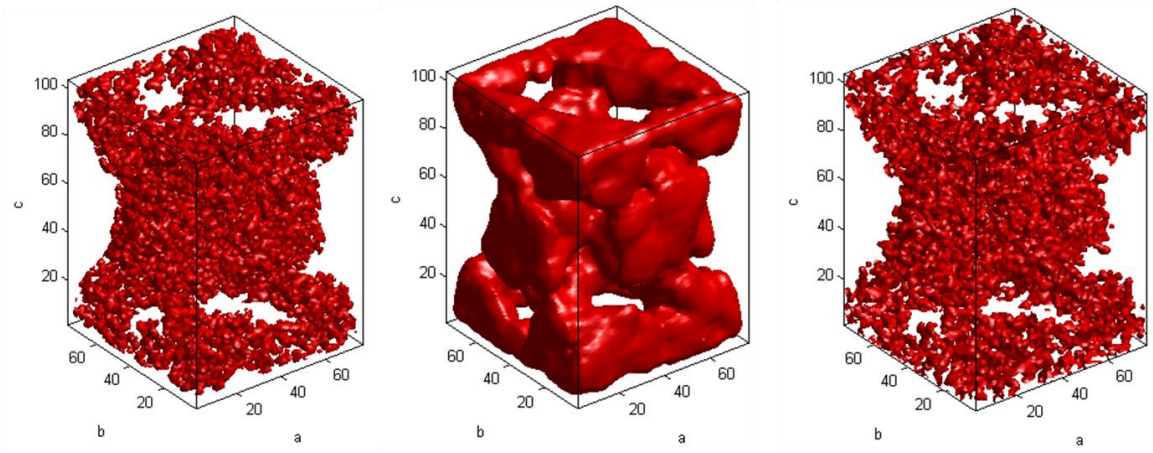


Figure 4.29 Charge density distribution of (a) Streptavidin with PEG and (c) Streptavidin free from ligand. (b) Envelope estimated from streptavidin free from ligand.

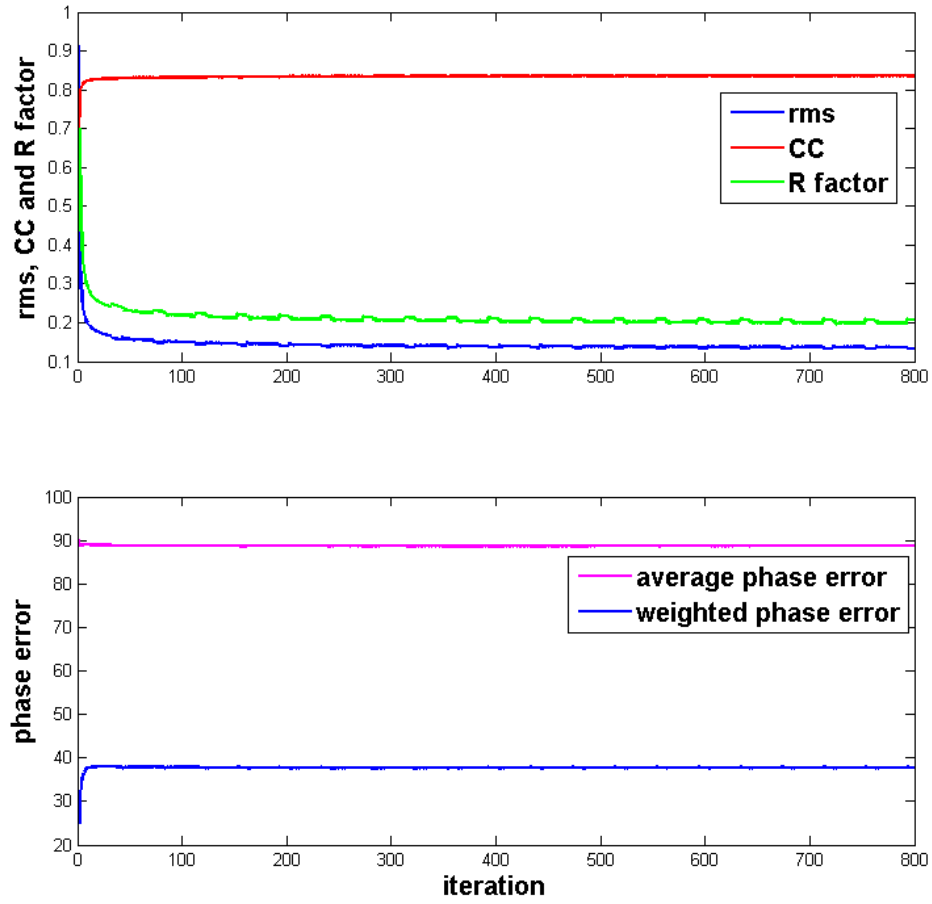
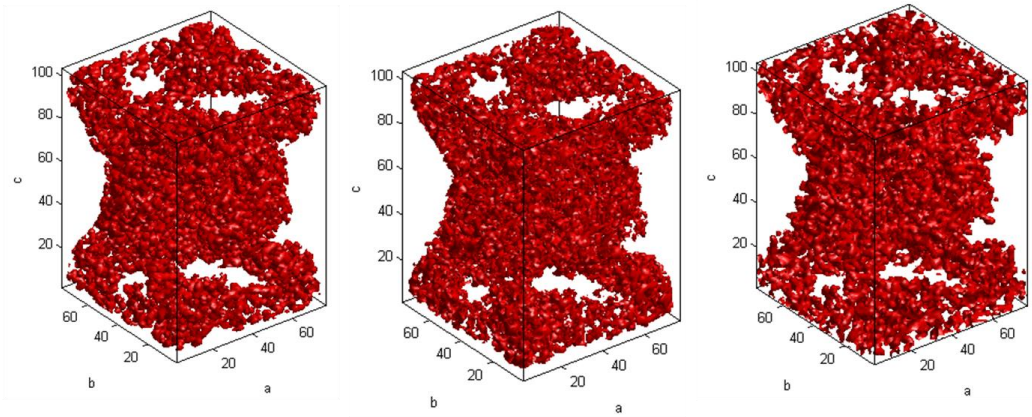


Figure 4.30 Correlation coefficient CC and rms value over iteration



(a) 3rdu (b) reconstruction (c) 3rdx

Figure 4.31 Comparison between (a) original model and (b) reconstruction. (c) Model used for generating support.

In figure 4.30, the CC values does not converge to 1 even with perfect structure factor amplitudes which are free from error. This is caused by the rough estimate using a model different from itself with many inconsistencies. Therefore there exists certain false constraints. This can be improved with known geometry shape from prior biology knowledge/Molecular replacement. This method actually imposes a much weaker constraint instead of directly taking phases from a model. An envelope is intrinsically a binary mask. The internal structure is reconstructed by HIO algorithm.

4.3.7 *Parameter optimization*

In this approach, the phases of a small number (up to 10) of low-order of reflections (and their symmetry-related mates) were treated as free parameters in the HIO optimization, and a search conducted over all possible values of these phases. Here we used rms, R factor as metric. We also used cc value, which is unavailable without a known model. We found that the lowest rms values are very close.

We started with treating phases associated with Bragg spots within 80 Å resolution as free parameters. There are 6 Bragg spots in total, namely $(100), (\bar{1}00), (010), (0\bar{1}0), (001), (00\bar{1})$. But only three are independent by symmetry constraints and Friedel's law. We sample phases $\varphi(100), \varphi(010), \varphi(001)$ from 10° to 360°, with 10° interval. Hence, there are 36^3 combination of initial phases. In the following simulations, these phases are additional constraints in reciprocal space, apart from the known intensities. The computational run time is about 50 hours. We found that rms value ranges from 0.0151 to 0.0169, and R factor ranges from 0.0292 to 0.0325. The minimum is reached at $(360^{\circ}, 170^{\circ}, 130^{\circ})$. When we use this optimal angle as initial constraints, we found the rms and R factor changes at the same number of iteration. Moreover, the structure is not reconstructed properly, as shown in following figure.

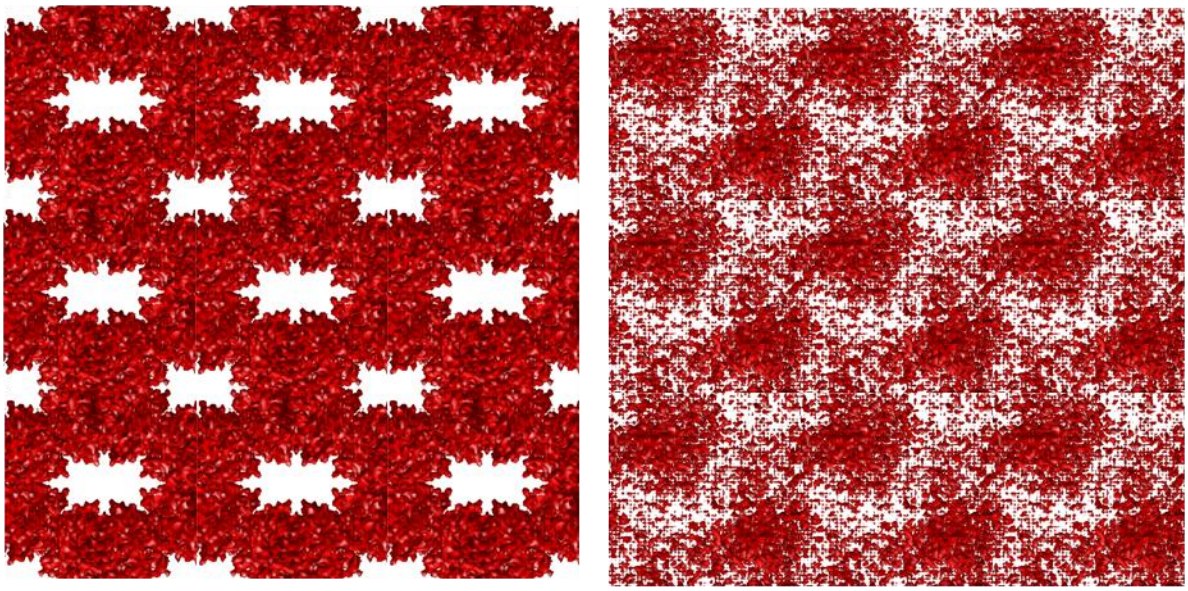


Figure 4.32 Charge density of model (a) and reconstruction (b) over several unit cells

This didn't work for several reasons. Firstly, different iterations may vary even given the same initial conditions, which makes it difficult to identify the optimal phase set. Secondly, there doesn't exist a good metric to pick out the best parameters without a model. We tried R factor and rms. We found, it may be higher for more known parameters. Thirdly, the HIO still couldn't improve much, even though a certain fraction of phases were known. The more phase we provide, the better structure we got at the convergence of HIO algorithm.

Provided with a sufficient number of known phases, a rough structure could be obtained at the convergence. We find the minimum $|\vec{k}|$ required for successful structure reconstruction is $0.025\ \text{Å}^{-1}$(10 independent phases considering symmetry and Friedel's law), with $CC = 0.748$, as shown in Fig 6. However the computational cost of this approach rapidly becomes prohibitive. Also, we need a better metric to pick out the optimal phase combinations. It is impossible to obtain structure by optimizing phasing both theoretically or computationally.

4.3.8 *Prior phases*

When phases associated with low resolution structure factor amplitudes are available from cryoEM or molecular replacement , it could also help HIO algorithm converge to a higher CC value. Here, we supply all the phases within radius $|\vec{k}|(\text{Å}^{-1})$ in reciprocal space as prior information. Figure 4.33 shows that CC converges to a higher value when more phases are supplied.

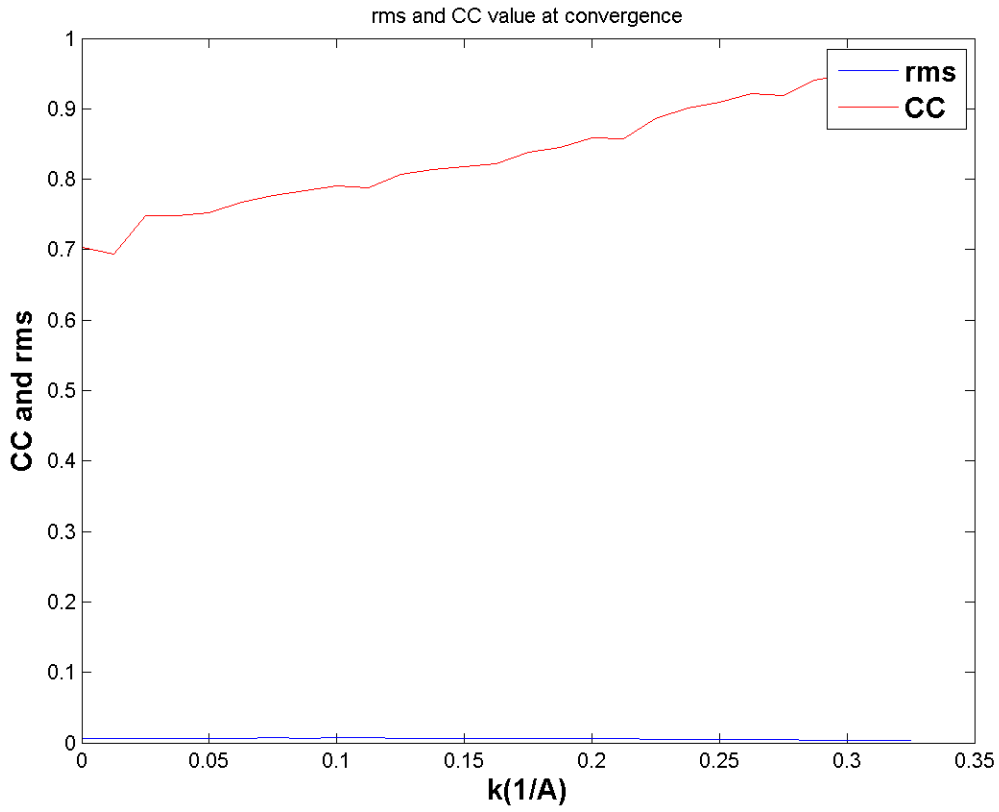


Figure 4.33 Correlation coefficient CC and rms value from HIO algorithm at convergence plotted against the radius of $|\vec{k}|$ vector in reciprocal space, within which known phases have been supplied to the algorithm.

Our simulation indicates that CC is the key value to evaluate whether the estimate structure is good or not. We find the minimum $|\vec{k}|$ required for successful structure reconstruction

is $0.025\ \text{Å}^{-1}$, with $CC = 0.748$, as shown in Fig 5 It means that phase could be retrieved for 3 Å diffraction data provided 40 Å resolution images at various orientations are available

4.4 Artificial 2D crystal

The main difficulty in achieving real ab-initio phasing is that we don't have lateral support. If we can make artificial 2D crystals with bigger space, then we may sample finer and have a lateral support which enable ab-inito phasing possible. Creating more space in between unit cells is the most straightforward to way to achieve ab-initio phasing (for example: creating a sample holder with some inorganic material to embed molecules). It's hard to create more space naturally as interaction will be too weak to make molecule organized by itself. In this way, the signal is still amplified, proportional to N^2.

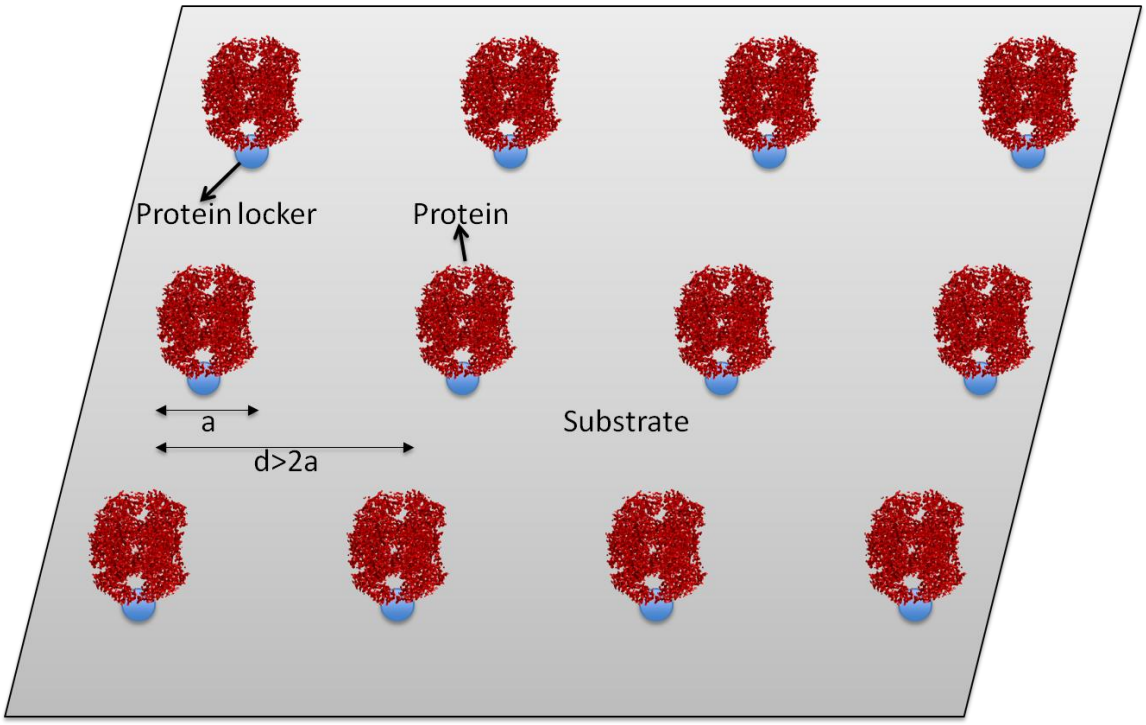
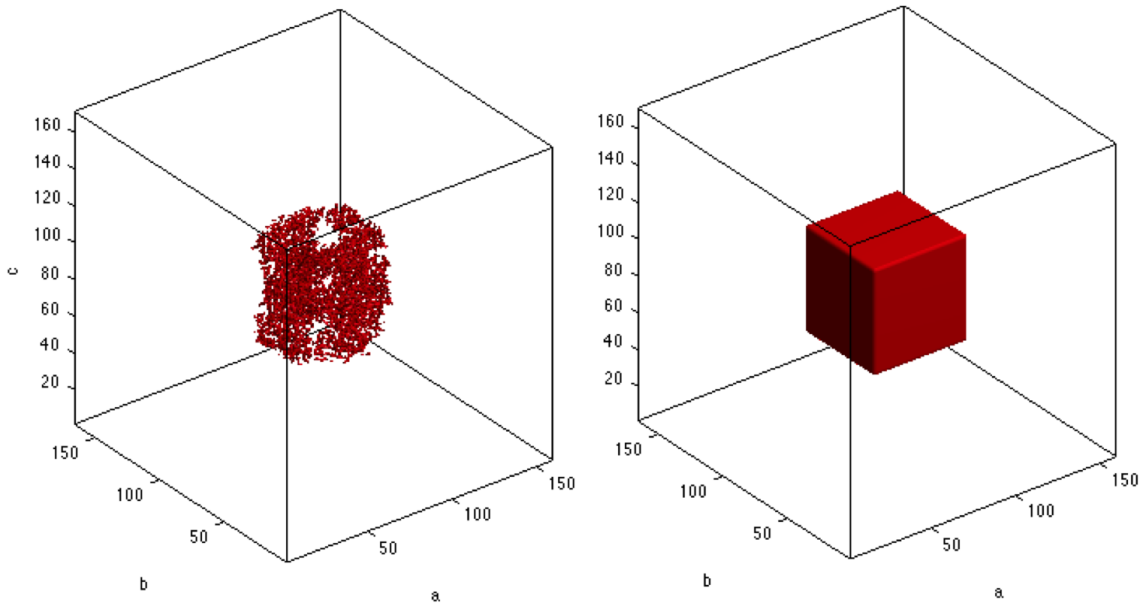


Figure 4.34 Artificial 2D crystal

To demonstrate the theoretical feasibility of phasing this system, here I take a single unit cell as a particle, with no periodicity in each dimension. Then the intensity distribution in reciprocal space is continuous in every direction. Hence a compact support can be applied in each side of the unit cell in real space.

In the following simulation, zeros are padded around the unit cell to generate a super cell, with a lattice constant three times bigger in each side, shown in fig 4.35. Then structure factor were calculated to 3 Å resolution by taking the Fourier Transform of the electron density of the super cell. Only the structure amplitude and compact support in real space are constraint applied, without any further phase prior information.

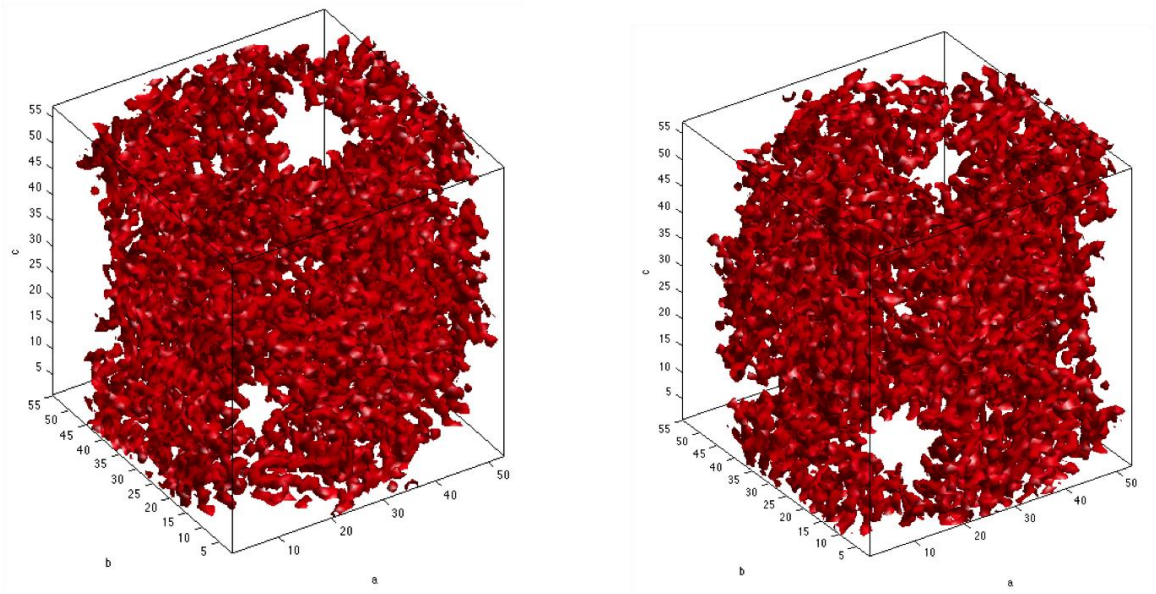


(a) Super cell with triple lattice constant in each side

(b) Support for super cell

Figure 4.35 Super cell with its support. The triple cell is shown at 8σ level in (a). Red pixels in (b) have value 1, while the empty space are zeros.

In this scenario, the structure recovered from HIO shares a high resemblance to the model, as shown in Fig 4.36. The CC and rms values are shown in Fig 4.37. It seems like there is an inversion relation between HIO estimate and our original model.



(a) 3A model

(b) HIO estimate

Figure 4.36 Comparison between model and HIO estimate in 3D view. Both are shown at 2σ level.

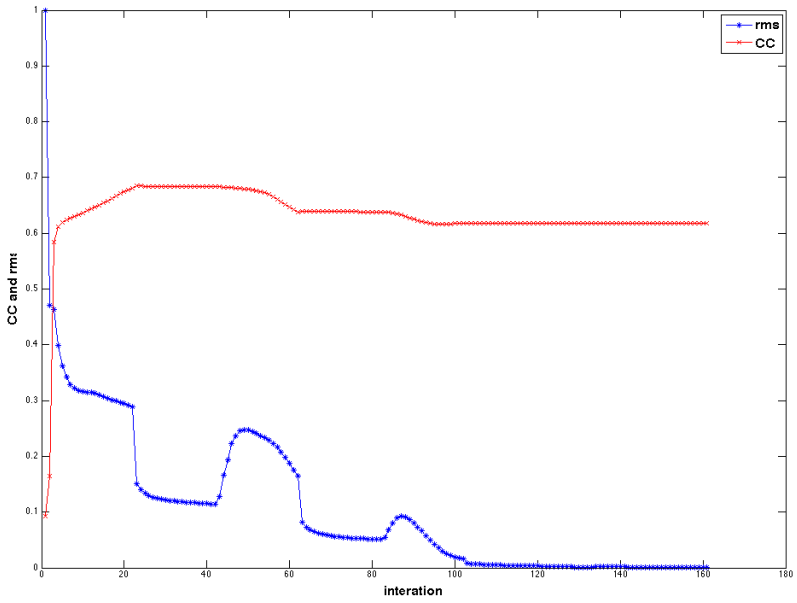


Figure 4.37 Correlation coefficient CC and rms value as a function of iterations, starting with known structure amplitudes and random phases

The size of a unit cell is typically in between 10~300 Å. If they are separated to sufficiently wide space on a substrate, then oversampling is possible. This technique may have advantages over single particle imaging in terms of signal strength and orientation control. The

sample holder is a mixture with (a) silicon and (b) a secondary material which binds the membrane proteins. Using a silicon holder should not cause any problems since the lattice constants for silicon are small in real space that the diffraction from the holder is very bright, sparse and can be predicted.

It is possible to make a slice of graphene and drill holes every 10nm. ASML EUV soft lithography device has 18nm resolution. Or we could put a "locker" to fix the protein at every 10 nm. In this way, we may grow 2D crystals very quick given such a substrate. This type of experiment can only be achieved at XFEL, since radiation damage would become a significant deterrent to study such samples. We benefit a lot from the new design. First, the signal is much stronger than single particle case. Ideally, the signal can be amplified by N^2. So, this experiment may be even conducted in a hydrated environment which may preserve it's functionality. Second, the orientation between crystal and X-ray beam can be recorded using a goniometer. It will relieve a lot of effort on data analysis. Third, this method is easy for mass production. The substrate and locker is the most crucial aspect of the experiment. If the secondary medium is identified, that can glue many proteins, a substantial amount of efforts and time in growing crystals would be reduced.

4.5 Conclusion and prospectus

"There ain't no such thing as a free lunch." The phase information is lost since detector can only record the magnitudes of complex structure factor. Phases can't be retrieved from nothing and it is not naturally inscribed in diffraction pattern from any system (single particle or crystal) without any prior knowledge. When "oversampling" is referred, an assumption has immediately been made that the object size is known. Otherwise, it would not be possible to estimate whether the diffraction pattern is oversampled or under-sampled. Even for single particle imaging, there is a key implicit information used for obtaining support - "single particle". It is known that charge density beyond a certain boundary will be zeros and this information is the key to obtaining support from autocorrelation function. This is also the reason why it is very unlikely to succeed in phasing diffraction data from crystals.

In this chapter, we demonstrated the application of iterative phasing algorithm in phasing two-dimensional crystal diffraction data. Structure can be retrieved only if a sufficiently tight support/molecular envelope is available. The size of the envelope is limited by the fraction of disordered volume per unit cell, which typically refers to solvent fraction. However, with certain ordered regions of the solvent and certain other regions being flexible, they need not be exactly the same. This is also the advantage of crystallography over EM imaging since the signal from the ordered region is greatly amplified, making it possible to distinguish the solvent molecule and protein by charge density.

The fact that the phase problem is hard to solve is largely due to two factors: 1) available data doesn't guarantee unique solution; 2) the unique solution exists, but there is no powerful algorithm to find it. Currently, several iterative projection algorithm variants are proposed to address this question. This thesis mainly addressed the first case with the standard Hybrid Input-Output algorithm, which is widely accepted and a successful algorithm in image processing. We found that as long as enough prior information is available, it can lead the algorithm towards the the right solution. This algorithm is also very convenient to integrate various experimental results in iteration. Therefore, it's worthwhile to develop HIO algorithm that can implement additional constraints with protein information from various experimental results, such as NMR, histogram matching etc.

Artificial two-dimensional crystal preserves the feasibility of ab-initio phasing and has a moderate signal level which is much stronger than from a single particle, but weaker than from a 3D crystal. The X-ray diffraction experiment can only be achieved at an XFEL since it's structure is unstable. If the substrate is easy to make, crystallization would be greatly simplified. The data analysis would be straightforward with iterative projection algorithms and It will open up a new field in X-ray crystallography.

REFERENCES

Aquila, A., Hunter, M. S., Doak, R. B., Kirian, R. a, Fromme, P., White, T. a, … Chapman, H. N. (2012). Time-resolved protein nanocrystallography using an X-ray free-electron laser. *Optics Express*, *20*(3), 2706–16. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23633593

Arnlund, D., Johansson, L. C., Wickstrand, C., Barty, A., Williams, G. J., Malmerberg, E., … Neutze, R. (2014). Visualizing a protein quake with time-resolved X-ray scattering at a free-electron laser. *Nature Methods*, *11*(9). http://doi.org/10.1038/nmeth.3067

Assumus, A. (1995). Early History of X Raus. *Beam Line*, 10–24. Retrieved from http://www.slac.stanford.edu/pubs/beamline/25/2/25-2-assmus.pdf

Barty, A., Kirian, R. a., Maia, F. R. N. C., Hantke, M., Yoon, C. H., White, T. a., & Chapman, H. (2014). Cheetah: Software for high-throughput reduction and analysis of serial femtosecond X-ray diffraction data. *Journal of Applied Crystallography*, *47*(3), 1118–1131. http://doi.org/10.1107/S1600576714007626

Bencharit, S. (2012). History of Progress and Challenges in Structural Biology. *J Pharmacogenom Pharmacoproteomics*, *S4*. http://doi.org/10.4172/2153-0645.S4-e001

Bogan, M. J., Benner, W. H., Boutet, S., Rohner, U., Frank, M., Barty, A., … Chapman, H. N. (2008). Single Particle X-ray Diffractive Imaging. *Nano Letters*, *8*(1), 310–316. http://doi.org/10.1021/nl072728k

Bogan, M. J., Boutet, S., Barty, A., Benner, W. H., Frank, M., Lomb, L., … Chapman, H. N. (2010). Single-shot femtosecond x-ray diffraction from randomly oriented ellipsoidal nanoparticles. *Physical Review Special Topics - Accelerators and Beams*, *13*(9), 94701. http://doi.org/10.1103/PhysRevSTAB.13.094701

Bogan, M. J., Starodub, D., Hampton, C. Y., & Sierra, R. G. (2010). Single-particle coherent diffractive imaging with a soft x-ray free electron laser: towards soot aerosol morphology. *Journal of Physics B: Atomic, Molecular and Optical Physics*, *43*(19), 194013. http://doi.org/10.1088/0953-4075/43/19/194013

Chapman, H. N., Barty, A., Marchesini, S., Noy, A., Hau-riege, S. P., Cui, C., … Shapiro, D. (2006). High-resolution ab initio three-dimensional x-ray diffraction microscopy, *23*(5).

Chapman, H. N., Fromme, P., Barty, A., White, T. a, Kirian, R. a, Aquila, A., … Spence, J. C. H. (2011). Femtosecond X-ray protein nanocrystallography. *Nature*, *470*(7332), 73–7. http://doi.org/10.1038/nature09750

Conrad, C. E., Basu, S., James, D., Wang, D., Schaffer, A., Roy-Chowdhury, S., … Fromme, P. (2015). A novel inert crystal delivery medium for serial femtosecond crystallography. *IUCrJ*, *2*, 421–430. http://doi.org/10.1107/S2052252515009811

Dempster, A. P., Laird, N. M., & B., R. D. (2007). Maximum Likelihood from Incomplete Data via the EM Algorithm A., *39*(1), 1–38.

Elser, V. (2009). Noise Limits on Reconstructing Diffraction Signals from Random Tomographs. *IEEE Trans. Inf. Theor.*, *55*(10), 4715–4722. article. http://doi.org/10.1109/TIT.2009.2027547

Elser, V. (2011). Three-dimensional structure from intensity correlations. *New J. Phys.*, *13*. http://doi.org/10.1088/1367-2630/13/12/123014 Abstract.

Fienup, J. R. (2004). Data Processing in Support of Image Reconstruction from X-Ray Diffraction Data of Nonperiodic Objects, (June), 1–16.

Fienup, J. R., Crimmins, T. R., & Holsztynski, W. (1982). Reconstruction of the support of an object from the support of its autocorrelation. *Journal of the Optical Society of America*, *72*(5), 610. http://doi.org/10.1364/JOSA.72.000610

Frank, M., Carlson, D. B., Hunter, M. S., Williams, G. J., Messerschmidt, M., Zatsepin, N. a., … Evans, J. E. (2014). Femtosecond X-ray diffraction from two-dimensional protein crystals. *IUCrJ*, *1*(2), 95–100. http://doi.org/10.1107/S2052252514001444

Fung, R., Shneerson, V., Saldin, D. K., & Ourmazd, A. (2008). Structure from fleeting illumination of faint spinning objects in flight. *Nature Physics*, *5*(1), 64–67. http://doi.org/10.1038/nphys1129

Hajdu, J. (2003). Diffraction imaging of single particles and biomolecules, *144*, 219–227. http://doi.org/10.1016/j.jsb.2003.09.025

Hart, P., Boutet, S., Carini, G., Dubrovin, M., Duda, B., Fritz, D., … Morse, J. (2012). The CSPAD megapixel x-ray camera at LCLS. *Proceedings of SPIE*, *8504*, 85040C–85040C–11. http://doi.org/10.1117/12.930924

Hartley, H. O. (1958). Maximum Likelihood Estimation from Incomplete Data. *Biometrics*, *14*(2), 174–194. JOUR. http://doi.org/10.2307/2527783

Hauptman, H. A. (1991). History of X-ray crystallography, *10*, 13–18. http://doi.org/10.1007/BF00674136

Hendrickson, W. A. (2013). Evolution of diffraction methods for solving crystal structures. *Acta Crystallographica Section A: Foundations of Crystallography*, *69*(1), 51–59. http://doi.org/10.1107/S0108767312050453

Hunter, M. S., Segelke, B., Messerschmidt, M., Williams, G. J., Zatsepin, N. a, Barty, A., … Frank, M. (2014). Fixed-target protein serial microcrystallography with an x-ray free electron laser. *Scientific Reports*, *4*, 6026. http://doi.org/10.1038/srep06026

Kam, Z. (1977). Determination of Macromolecular Structure in Solution by Spatial Correlation of Scattering Fluctuations. *Macromolecules*, *10*(5), 927–934. JOUR. http://doi.org/10.1021/ma60059a009

Kam, Z. (1980). The reconstruction of structure from electron micrographs of randomly oriented particles. *Journal of Theoretical Biology*, *82*(1), 15–39. JOUR. http://doi.org/http://dx.doi.org/10.1016/0022-5193(80)90088-0

Kirian, R. a. (2012). Structure determination through correlated fluctuations in x-ray scattering. *Journal of Physics B: Atomic, Molecular and Optical Physics*, *45*(22), 223001. http://doi.org/10.1088/0953-4075/45/22/223001

Kirian, R. A., Awel, S., Eckerskorn, N., Fleckenstein, H., Weidorn, M., Adriano, L., … Chapman, H. N. (2015). Simple convergent-nozzle aerosol injector for single-particle diffractive imaging with X-ray free-electron lasers. *Structural Dynamics*, *2*, 41717. http://doi.org/10.1063/1.4922648

Kirian, R. a, Schmidt, K. E., Wang, X., Doak, R. B., & Spence, J. C. H. (2011). Signal, noise, and

resolution in correlated fluctuations from snapshot small-angle x-ray scattering. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, *84*(1–1), 11921. http://doi.org/10.1103/PhysRevE.84.011921

Kirian, R. a, Wang, X., Weierstall, U., Schmidt, K. E., Spence, J. C. H., Hunter, M., … Holton, J. (2010). Femtosecond protein nanocrystallography-data analysis methods. *Optics Express*, *18*(6), 5713–23. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20389587

Kupitz, C., Basu, S., Grotjohann, I., Fromme, R., Zatsepin, N. A., Rendek, K. N., … Fromme, P. (2014). Serial time-resolved crystallography of photosystem II using a femtosecond X-ray laser. *Nature*, *513*(7517), 261–5. http://doi.org/10.1038/nature13453

Liu, W., Wacker, D., Gati, C., Han, G. W., James, D., Wang, D., … Cherezov, V. (2013). Serial Femtosecond Crystallography of G Protein-Coupled Receptors. *Science (New York, N.Y.)*, *342*(6165), 1521–1524. http://doi.org/10.1126/science.1244142.Serial

Liu, W., Wacker, D., Gati, C., Han, G. W., James, D., Wang, D., … Cherezov, V. (2013). Serial Femtosecond Crystallography of G Protein-Coupled Receptors in Lipidic Cubic Phase. *Science (New York, N.Y.)*.

Loh, N. D. (2013). No Title. Retrieved from http://www.duaneloh.com/

Loh, N. D., Bogan, M. J., Elser, V., Barty, A., Boutet, S., Bajt, S., … Chapman, H. N. (2000). Cryptotomography: reconstructing 3D Fourier intensities from randomly oriented single-shot diffraction patterns, 5–8.

Loh, N.-T. D., & Elser, V. (2009). Reconstruction algorithm for single-particle diffraction imaging experiments. *Physical Review E*, *80*(2), 26705. JOUR. Retrieved from http://link.aps.org/doi/10.1103/PhysRevE.80.026705

Magalhães, M. L. B., Czekster, C. M., Guan, R., Malashkevich, V. N., Almo, S. C., & Levy, M. (2011). Evolved streptavidin mutants reveal key role of loop residue in high-affinity binding. *Protein Science*, *20*(7), 1145–1154. http://doi.org/10.1002/pro.642

Mclachlan, G. J., & Krishnan, T. (1977). *The EM Algorithm and Extensions Second Edition*. John Wiley & Sons.

Miao, J., Charalambous, P., Kirz, J., & Sayre, D. (1999). Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, *400*(6742), 342–344. http://doi.org/10.1038/22498

Miao, J., Ishikawa, T., Robinson, I. K., & Murnane, M. M. (2015). Beyond crystallography: Diffractive imaging using coherent x-ray light sources. *Science*, *348*(6234), 530–535. http://doi.org/10.1126/science.aab0097

Miao, J., Kirz, J., & Sayre, D. (2000). The oversampling phasing method. *Acta Crystallographica Section D: Biological Crystallography*, *56*(10), 1312–1315. http://doi.org/10.1107/S0907444900008970

Miao, J., Sayre, D., & Chapman, H. N. (1998). Phase retrieval from the magnitude of the Fourier transforms of nonperiodic objects. *Journal of the Optical Society of America A*. http://doi.org/10.1364/JOSAA.15.001662

Narumi, S., & Sautter, H. (2011). SACLA X-ray Free Electron Laser Facility: Shortest Wavelength Ever Brings Us Closer to the World of Atoms. Retrieved from

http://www.nippon.com/en/features/c00501/

Neutze, R., Wouts, R., van der Spoel, D., Weckert, E., & Hajdu, J. (2000). Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature*, *406*(6797), 752–7. http://doi.org/10.1038/35021099

Pedrini, B., Tsai, C., Capitani, G., Padeste, C., Hunter, M. S., Zatsepin, N. A., … Schertler, G. F. X. (2014). diffraction at Linac Coherent Light Source ˚ resolution in protein two- dimensional-crystal X-ray diffraction at Linac Coherent Light Source.

Pellegrini, C., & Stöhr, J. (2009). X-Ray Free Electron Lasers: Principles, Properties and Applications, 1–16. http://doi.org/10.1016/S0168-9002(03)00739-3

Philipp, H. T., Ayyer, K., Tate, M. W., Elser, V., & Gruner, S. M. (2012). Solving structure with sparse, randomly-oriented x-ray data. *Optics Express*, *20*(12), 13129. http://doi.org/10.1364/OE.20.013129

Rossmann, M. (1990). The molecular replacement method. *Acta Crystallographica Section A: Foundations of …*, 73–82. Retrieved from http://scripts.iucr.org/cgi-bin/paper?s0108767389009815

Rossmann, M. G., & Blow, D. M. (1962). The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallographica*, *15*(1), 24–31. http://doi.org/10.1107/S0365110X62000067

Saldin, D. K., Poon, H. C., Bogan, M. J., Marchesini, S., Shapiro, D. a., Kirian, R. a., … Spence, J. C. H. (2011). New Light on Disordered Ensembles: Ab Initio Structure Determination of One Particle from Scattering Fluctuations of Many Copies. *Physical Review Letters*, *106*(11), 115501. http://doi.org/10.1103/PhysRevLett.106.115501

Saldin, D. K., Poon, H. C., Shneerson, V. L., Howells, M., Chapman, H. N., Kirian, R. A., … Spence, J. C. H. (2010). Beyond small-angle x-ray scattering: Exploiting angular correlations. *Physical Review B*, *81*(17), 174105. JOUR. Retrieved from http://link.aps.org/doi/10.1103/PhysRevB.81.174105

Saldin, D. K., Poon, H.-C., Schwander, P., Uddin, M., & Schmidt, M. (2011). Reconstructing an icosahedral virus from single-particle diffraction experiments. *Optics Express*, *19*(18), 17318–35. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21935096

Saldin, D. K., & Shneerson, V. L. (n.d.). Structure of a single particle from scattering by many particles randomly oriented about an axis : towards structure solution without crystallization ?

Sauter, N. K., Hattne, J., Grosse-Kunstleve, R. W., & Echols, N. (2013). New Python-based methods for data processing. *Acta Crystallographica. Section D, Biological Crystallography*, *69*(Pt 7), 1274–82. http://doi.org/10.1107/S0907444913000863

Scardi, P., Mccusker, L. B., Dreele, R. B. Von, Cox, D. E., & Loue, D. (1999). Rietveld refinement guidelines, 36–50. http://doi.org/10.1107/S0021889898009856

Schlichting, I., & Miao, J. (2012). Emerging opportunities in structural biology with X-ray free-electron lasers. *Current Opinion in Structural Biology*, *22*(5), 613–626. http://doi.org/10.1016/j.sbi.2012.07.015

Seibert, M. M., Ekeberg, T., Maia, F. R. N. C., Svenda, M., Andreasson, J., Jönsson, O., … Hajdu, J. (2011). Single mimivirus particles intercepted and imaged with an X-ray laser. *Nature*,

470(7332), 78–81. http://doi.org/10.1038/nature09748

Shapiro, D. A., Chapman, H. N., DePonte, D., Doak, R. B., Fromme, P., Hembree, G., … Weierstall, U. (2008). Powder diffraction from a continuous microjet of submicrometer protein crystals. *Journal of Synchrotron Radiation*, *15*(6), 593–599. article. http://doi.org/10.1107/S0909049508024151

Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J., & Segev, M. (2015). Phase Retrieval with Application to Optical Imaging: A contemporary overview. *IEEE Signal Processing Magazine*, *32*(3), 87–109. http://doi.org/10.1109/MSP.2014.2352673

Spence, J. C. H., Weierstall, U., & Chapman, H. N. (2012). X-ray lasers for structural and dynamic biology. *Reports on Progress in Physics. Physical Society (Great Britain)*, *75*(10), 102601. http://doi.org/10.1088/0034-4885/75/10/102601

Spence, J. C. H., Weierstall, U., Fricke, T. T., Glaeser, R. M., & Downing, K. H. (2003). Three-dimensional diffractive imaging for crystalline monolayers with one-dimensional compact support, *144*, 209–218. http://doi.org/10.1016/j.jsb.2003.09.019

Weierstall, U., James, D., Wang, C., White, T. a, Wang, D., Liu, W., … Cherezov, V. (2014). Lipidic cubic phase injector facilitates membrane protein serial femtosecond crystallography. *Nature Communications*, *5*, 3309. http://doi.org/10.1038/ncomms4309

Weierstall, U., Spence, J. C. H., & Doak, R. B. (2012). Injector for scattering measurements on fully solvated biospecies. *Review of Scientific Instruments*, *83*(3), 35108. http://doi.org/10.1063/1.3693040

White, T. a., Kirian, R. a., Martin, A. V., Aquila, A., Nass, K., Barty, A., & Chapman, H. N. (2012). CrystFEL: A software suite for snapshot serial crystallography. *Journal of Applied Crystallography*, *45*(2), 335–341. http://doi.org/10.1107/S0021889812002312

White, T. a, Barty, A., Stellato, F., Holton, J. M., Kirian, R. a, Zatsepin, N. a, & Chapman, H. N. (2013). Crystallographic data processing for free-electron laser sources. *Acta Crystallographica. Section D, Biological Crystallography*, *69*(Pt 7), 1231–40. http://doi.org/10.1107/S0907444913013620

APPENDIX A

NOTATIONS IN ANGULAR CORRELATION FUNCTION ALGORITHM

$I(\mathbf{q}_l, \omega_k^l)$ the intensity contribution from the l_th crystal during k_th snapshot, with its orientation ω_k^l, scattering vector $\mathbf{q}_l$ (Bold letter means a vector, letters without bold mean magnitude).

$I_k(\mathbf{q}_l)$ the observed intensity from k_th snapshot

Note: In the following of this report, I without subscript k always means the diffraction intensity from one crystal. I_k always represents the observed diffraction intensity which results from x-ray scattered by many crystals.

$\tilde{I}_k(\mathbf{q}_l)$ fluctuation intensity from k_th snapshot

$C_1(\mathbf{q}_l, \mathbf{q}_j, \Delta\varphi)$ angular pair correlation function for single crystal

$T_1(\mathbf{q}_l, \mathbf{q}_j, \Delta\varphi)$ angular triple correlation function for single crystal

$\tilde{C}_{exp}(\mathbf{q}_l, \mathbf{q}_j, \Delta\varphi)$ fluctuation angular pair correlation function for multiple crystals, which is averaged over all experimental or simulated powder diffraction patterns.

$\tilde{T}_{exp}(\mathbf{q}_l, \mathbf{q}_j, \Delta\varphi)$ fluctuation angular triple correlation function for multiple crystals, which is averaged over all experimental or simulated powder diffraction patterns.

$B_m(\mathbf{q}_l, \mathbf{q}_j)$ Fourier transform of $C_1(\mathbf{q}_l, \mathbf{q}_j, \Delta\varphi)$

$FT(\mathbf{q}_l, \mathbf{q}_j, \Delta\varphi)$ Fourier transform of $T_1(\mathbf{q}_l, \mathbf{q}_j, \Delta\varphi)$

APPENDIX B

PROOF OF ANGULAR CORRELATION FUNCTION RETRIEVAL

For the diffraction pattern from a single crystal, the pair correlation function for two different rings is defined as

$$C_1(q_i, q_j, \Delta\varphi) = \frac{1}{N_\varphi}\sum_{m}^{N_\varphi} I(q_i, \varphi_m) I(q_j, \varphi_m + \Delta\varphi) \qquad (1)$$

where q_i and q_j represents radius of the i-th and j-th ring on diffraction pattern.. N_φ is the number of azimuthal angels at φ_m which the intensity are measured. In a similar way, the triple correlation function is defined as

$$T_1(q_i, q_j, \Delta\varphi) = \frac{1}{N_\varphi}\sum_{m}^{N_\varphi} I(q_i, \varphi_m)^2 I(q_j, \varphi_m + \Delta\varphi) \qquad (2)$$

Now let's consider many-crystal case. Here we assume each crystal scatters x-ray incoherently, thus the intensity observed on detector is simply the sum of the intensity from each individual crystal.

$$I_k(\mathbf{q_1}) = \sum_{l}^{N_c} I(\mathbf{q_1}, \omega_k^l)$$

where ω_k^l is the orientation of l_th crystal during k_th snapshot.

The fluctuation intensity is defined as

$$\delta I_k(\mathbf{q_1}) = I_k(\mathbf{q_1}) - < I_k(\mathbf{q_1}) >_k$$

where the second term means average over all diffraction patterns

$$< I_k(\mathbf{q_1}) >_k = \frac{1}{N_d}\sum_{k=1}^{N_d} I_k(\mathbf{q_1})$$

where N_d is the total number of diffraction patterns.

The fluctuation pair correlation for simulated diffraction patterns is defined as

$$\delta C_{exp}(q_i, q_j, \Delta\varphi) = \frac{1}{N_d}\sum_{k}^{N_d} \frac{1}{2\pi}\int_0^{2\pi} \delta I_k(q_i, \varphi)\delta I_k(q_j, \varphi + \Delta\varphi)\, d\varphi$$

Let's change the integral by sum,

$$\check{C}_{exp}(q_i, q_j, \Delta\varphi) = \frac{1}{N_\varphi}\frac{1}{N_d}\sum_m^{N_\varphi}\sum_k^{N_d} \check{I}_k(q_i, \varphi_m)\check{I}_k(q_j, \varphi_m + \Delta\varphi)$$

$$= \frac{1}{N_\varphi}\frac{1}{N_d}\sum_m^{N_\varphi}\sum_k^{N_d}\left(\sum_l^{N_c} I\left(q_i, \varphi_m, \omega_k^l\right) - < I_k(q_i) >_k\right)\left(\sum_n^{N_c} I\left(q_j, \varphi_m + \Delta\varphi, \omega_k^n\right) - < I_k(q_j) >_k\right)$$

$$= \frac{1}{N_\varphi}\frac{1}{N_d}\sum_m^{N_\varphi}\sum_k^{N_d}\left(\sum_l^{N_c} I\left(q_i, \varphi_m, \omega_k^l\right)\sum_n^{N_c} I\left(q_j, \varphi_m + \Delta\varphi, \omega_k^n\right) - < I_k(q_i) >_k \sum_n^{N_c} I\left(q_j, \varphi_m + \Delta\varphi, \omega_k^m\right)\right.$$
$$\left. - < I_k(q_j) >_k \sum_l^{N_c} I\left(q_i, \varphi_m, \omega_k^l\right) + < I_k(q_i) >_k < I_k(q_j) >_k\right)$$

$$= \frac{1}{N_\varphi}\frac{1}{N_d}\sum_m^{N_\varphi}\sum_k^{N_d}\left(\sum_{l=n}^{N_c} I\left(q_i, \varphi_m, \omega_k^l\right) I\left(q_j, \varphi_m + \Delta\varphi, \omega_k^l\right) + \sum_{l\neq n}^{N_c} I\left(q_i, \varphi_m, \omega_k^l\right) I\left(q_j, \varphi_m + \Delta\varphi, \omega_k^n\right)\right.$$
$$- < I_k(q_i) >_k \sum_l^{N_c} I\left(q_j, \varphi_m + \Delta\varphi, \omega_k^l\right) - < I_k(q_j) >_k \left.\sum_l^{N_c} I\left(q_i, \varphi_m, \omega_k^l\right)\right)$$
$$+ < I_k(q_i) >_k < I_k(q_j) >_k$$

Note that in the first term, $\frac{1}{N_\varphi}\sum_m^{N_\varphi} I\left(q_i, \varphi_m, \omega_k^l\right) I\left(q_j, \varphi_m + \Delta\varphi, \omega_k^l\right)$, is the pair correlation function $C_1(q_i, q_j, \Delta\varphi)$ that would arise from single crystal. The second uncorrelated term can be expressed as

$$\sum_{l\neq n}^{N_c}\left(\frac{1}{N_\varphi}\frac{1}{N_d}\sum_m^{N_\varphi}\sum_k^{N_d} I\left(q_i, \varphi_m, \omega_k^l\right)\right) * \left(\frac{1}{N_\varphi}\frac{1}{N_d}\sum_m^{N_\varphi}\sum_k^{N_d} I\left(q_j, \varphi_m + \Delta\varphi, \omega_k^n\right)\right)$$

When we take average over all diffraction pattern and integrate over each ring, the relative orientation between crystals ω will be washed out. And both terms above will not depend on angle or any specific crystal. Here I simply denoted them as $\langle I(q_i)\rangle_\omega$ and $\langle I(q_j)\rangle_\omega$. As the sum $\sum_{l\neq n}^{N_c}$ has $(N_c{}^2 - N_c)$ terms. Hence, the second term has the following simple expression

$$= (N_c{}^2 - N_c)\langle I(q_i)\rangle_\omega \langle I(q_j)\rangle_\omega$$

And

$$\check{C}_{exp}(q_i, q_j, \Delta\varphi) = N_c C_1(q_i, q_j, \Delta\varphi) + (N_c{}^2 - N_c)\langle I(q_i)\rangle_\omega \langle I(q_j)\rangle_\omega - N_c < I_k(q_i) >_k \langle I(q_j)\rangle_\omega$$
$$- N_c \langle I(q_i)\rangle_\omega < I_k(q_j) >_k + < I_k(q_i) >_k < I_k(q_j) >_k$$

Note that $\langle I(q)\rangle_\omega$ is the average intensity from single crystal, $< I_k(q) >_k$ is the average intensity from N_c crystals during a x-ray shot. And $\langle I(q)\rangle_\omega$, $< I_k(q) >_k$ are uniform on each ring. So $\langle I(q)\rangle_\omega = \frac{1}{N_c} < I_k(q) >_k$. Then

$$\tilde{C}_{exp}(q_i, q_j, \Delta\varphi) = N_c C_1(q_i, q_j, \Delta\varphi) + (1 - \frac{1}{N_c} - 1 - 1 + 1) < I_k(q_i) >_k < I_k(q_j) >_k$$

Hence,

$$C_1(q_i, q_j, \Delta\varphi) = \frac{1}{N_c}\tilde{C}_{exp}(q_i, q_j, \Delta\varphi) + \frac{1}{N_c{}^2} < I_k(q_i) >_k < I_k(q_j) >_k$$

In the above equation, $\tilde{C}_{exp}(q_i, q_j, \Delta\varphi)$ and $< I_k(q_i) >_k$ can be easily calculated from diffraction patterns by definition. Hence, pair correlation for single crystal $C_1(q_i, q_j, \Delta\varphi)$ is solved.

APPENDIX C

DERIVATION FOR TRIPLE ANGULAR CORRELATION

The fluctuation triple correlation for simulated diffraction patterns is defined as

$$\tilde{T}_{exp}(q_i, q_j, \Delta\varphi) = \frac{1}{N_\varphi}\frac{1}{N_d}\sum_{m}^{N_\varphi}\sum_{k}^{N_d} \tilde{I}_k{}^2(q_i, \varphi_m)\tilde{I}_k(q_j, \varphi_m + \Delta\varphi)$$

This term can be directly calculated from all the diffraction patterns. In a similar fashion,

$\tilde{T}_{exp}(q_i, q_j, \Delta\varphi)$ can be expanded as

$$= \frac{1}{N_\varphi}\frac{1}{N_d}\sum_{m}^{N_\varphi}\sum_{k}^{N_d}\left(\left(\sum_{l}^{N_c} I_k(q_i, \omega_k^l)\right)^2 - 2 * \sum_{l}^{N_c} I_k(q_i, \omega_k^l) < I_k(q_i) >_k + < I_k(q_i) >_k{}^2\right)\left(\sum_{n}^{N_c} I_k(q_j, \omega_k^n) - < I_k(q_j) >_k\right)$$

$$= N_c T_1(q_i, q_j, \Delta\varphi) + (N_c{}^2 - N_c)\langle I_k(q_i)^2\rangle_\omega \langle I_k(q_j)\rangle_\omega - 2N_c\langle I_k(q_i)\rangle_k C_1(q_i, q_j, \Delta\varphi)$$
$$-(N_c{}^2 - N_c)\langle I_k(q_i)\rangle_k \langle I_k(q_i)\rangle_\omega \langle I_k(q_j)\rangle_\omega - N_c{}^2\langle I_k(q_i)\rangle_k{}^2\langle I_k(q_j, \omega_k^n)\rangle_\omega - N_c{}^2\langle I_k(q_i)^2\rangle_\omega \langle I_k(q_j, \omega_k^n)\rangle_k$$
$$+ 2N_c{}^2\langle I_k(q_i)\rangle_k \langle I_k(q_j)\rangle_k \langle I_k(q_i)\rangle_\omega + N_c{}^2\langle I_k(q_i)\rangle_k{}^2\langle I_k(q_j)\rangle_k$$
$$= N_c T_1(q_i, q_j, \Delta\varphi) - 2N_c\langle I_k(q_i)\rangle_k C_1(q_i, q_j, \Delta\varphi) + [N_c{}^2 - N_c - 2(N_c{}^2 - N_c) - N_c{}^2 - N_c{}^2 + 2N_c{}^2$$
$$+ N_c{}^2]\langle I_k(q_i)\rangle_k{}^2\langle I_k(q_j)\rangle_k$$
$$= N_c T_1(q_i, q_j, \Delta\varphi) - 2N_c\langle I_k(q_i)\rangle_k C_1(q_i, q_j, \Delta\varphi) + N_c\langle I_k(q_i)\rangle_k{}^2\langle I_k(q_j)\rangle_k$$

So

$$T_1(q_i, q_j, \Delta\varphi) = \frac{1}{N_c}\tilde{T}_{exp}(q_i, q_j, \Delta\varphi) + \frac{2}{N_c}\langle I_k(q_j)\rangle_k C_1(q_i, q_j, \Delta\varphi) - \frac{1}{N_c{}^3}\langle I_k(q_i)\rangle_k{}^2\langle I_k(q_j)\rangle_k$$

All terms on the right side of equation can be calculated from diffraction patterns, hence triple correlation for single crystal is solved.

APPENDIX D

FOURIER TRANSFORM OF PAIR ANGULAR CORRELATION

By definition, the Fourier transform of pair angular correlation is given by

$$B_m(q_i, q_j) = \frac{1}{2\pi}\int_0^{2\pi} C_1(q_i, q_j, \Delta\varphi)\exp(-im\Delta\varphi)\, d\Delta\varphi \qquad (1)$$

where

$$C_1(q_i, q_j, \Delta\varphi) = \frac{1}{2\pi}\int_0^{2\pi} I(q_i, \varphi_m) I(q_j, \varphi_m + \Delta\varphi)\, d\varphi$$

Expand $I(q, \varphi)$ in circular harmonics, then

$$C_1(q_i, q_j, \Delta\varphi) = \frac{1}{2\pi}\int_0^{2\pi} \left(\sum_n I_n(q_i)\exp(in\varphi)\right)\left(\sum_l I_l(q_j)\exp[il(\varphi + \Delta\varphi)]\right) d\varphi$$

$$= \frac{1}{2\pi}\int_0^{2\pi} \sum_n \sum_l I_n(q_i) I_l(q_j)\exp(in\varphi)\exp[il(\varphi + \Delta\varphi)]\, d\varphi \qquad (2)$$

Put eqn (2) to (1)

$$B_m(q_i, q_j) = \left(\frac{1}{2\pi}\right)^2 \int_0^{2\pi}\int_0^{2\pi} \sum_n \sum_l I_n(q_i) I_l(q_j)\exp(in\varphi)\exp[il(\varphi + \Delta\varphi)]\exp(-im\Delta\varphi)\, d\varphi\, d\Delta\varphi$$

$$= \left(\frac{1}{2\pi}\right)^2 \int_0^{2\pi}\int_0^{2\pi} \sum_n \sum_l I_n(q_i) I_l(q_j)\exp[i(n + l)\varphi]\exp[i(l - m)\Delta\varphi]\, d\varphi\, d\Delta\varphi$$

Note that

$$\frac{1}{2\pi}\int_0^{2\pi} \exp[i(l - m)\Delta\varphi]\, d\Delta\varphi = \begin{cases} 1 & \text{when } l = m \\ 0 & \text{when } l \neq m \end{cases}$$

Hence

$$B_m(q_i, q_j) = \frac{1}{2\pi}\int_0^{2\pi} \sum_n I_n(q_i) I_m(q_j)\exp[i(n + m)\varphi]\, d\varphi$$

Also note that

$$\frac{1}{2\pi}\int_0^{2\pi} \exp[i(n + m)\varphi]\, d\varphi = \begin{cases} 1 & \text{when } n = -m \\ 0 & \text{when } n \neq -m \end{cases}$$

So

$$B_m(q_i, q_j) = I_{-m}(q_i) I_m(q_j)$$

As $I_{-m}(q_i) = I_m^*(q_i)$, so

$$B_m(q_i, q_j) = I_m^*(q_i) I_m(q_j)$$

APPENDIX E

FOURIER TRANSFOMR OF TRIPLE ANGULAR CORRELATION

By definition, the Fourier transform of triple angular correlation is given by

$$FT_m^{(obs)}(q_i, q_j) = \frac{1}{2\pi}\int_0^{2\pi} T_1(q_i, q_j, \Delta\varphi)\exp(-im\Delta\varphi)\, d\Delta\varphi$$

or

$$FT_m^{(obs)}(q_i, q_j) = \frac{1}{N_\varphi}\sum_{\Delta\varphi} T_1(q_i, q_j, \Delta\varphi)\exp(-im\Delta\varphi)$$

where

$$T_1(q_i, q_j, \Delta\varphi) = \frac{1}{2\pi}\int_0^{2\pi} I(q_i, \varphi_m)^2 I(q_j, \varphi_m + \Delta\varphi)\, d\varphi$$

Expand $I(q, \varphi)$ in circular harmonics, then

$$T_1(q_i, q_j, \Delta\varphi) = \frac{1}{2\pi}\int_0^{2\pi}\left(\sum_n I_n(q_i)\exp(in\varphi)\right)^2\left(\sum_l I_l(q_j)\exp[il(\varphi + \Delta\varphi)]\right)d\varphi$$

$$= \frac{1}{2\pi}\int_0^{2\pi}\sum_{n_1}\sum_{n_2}\sum_l I_{n_1}(q_i)I_{n_2}(q_i)I_l(q_j)\exp(in_1\varphi)\exp(in_2\varphi)\exp[il(\varphi + \Delta\varphi)]\, d\varphi \qquad (2)$$

Put eqn (2) to (1)

$$FT_m(q_i, q_j) = \left(\frac{1}{2\pi}\right)^2\int_0^{2\pi}\int_0^{2\pi}\sum_{n_1}\sum_{n_2}\sum_l I_{n_1}(q_i)I_{n_2}(q_i)I_l(q_j)\exp(in_1\varphi)\exp(in_2\varphi)\exp[il(\varphi + \Delta\varphi)]\exp(-im\Delta\varphi)\, d\varphi\, d\Delta\varphi$$

$$= \left(\frac{1}{2\pi}\right)^2\int_0^{2\pi}\int_0^{2\pi}\sum_{n_1}\sum_{n_2}\sum_l I_{n_1}(q_i)I_{n_2}(q_i)I_l(q_j)\exp[i(n_1 + n_2 + l)\varphi]\exp[i(l - m)\Delta\varphi]\, d\varphi\, d\Delta\varphi$$

Note that

$$\frac{1}{2\pi}\int_0^{2\pi}\exp[i(l - m)\Delta\varphi]\, d\Delta\varphi = \begin{cases} 1 & \text{when } l = m \\ 0 & \text{when } l \neq m \end{cases}$$

Hence

$$FT_m(q_i, q_j) = \frac{1}{2\pi}\int_0^{2\pi}\sum_{n_1}\sum_{n_2} I_{n_1}(q_i)I_{n_2}(q_i)I_m(q_j)\exp[i(n_1 + n_2 + m)\varphi]\, d\varphi$$

Also note that

$$\frac{1}{2\pi}\int_0^{2\pi}\exp[i((n_1 + n_2 + m)\varphi]\, d\varphi = \begin{cases} 1 & \text{when } n_1 + n_2 = -m \\ 0 & \text{when } n_1 + n_2 \neq -m \end{cases}$$

So

$$FT_m(q_i, q_j) = \sum_{n_1} I_{n_1}(q_i)I_{-m-n_1}(q_i)I_m(q_j)$$

Let $n_1 = -M$, where $M = \pm 1, \pm 2, \cdots, \pm m_{max}$, except $M = m$

$$FT_m(q_i, q_j) = \sum_M I_{-M}(q_i)I_{M-m}(q_i)I_m(q_j)$$