December 2016

Symmetry and Reconstruction of Particle Structure from Random Angle Diffraction Patterns

Sandi Wibowo
University of Wisconsin-Milwaukee

SYMMETRY AND RECONSTRUCTION OF PARTICLE STRUCTURE FROM RANDOM ANGLE DIFFRACTION PATTERNS

by

Sandi Wibowo

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Physics

at

The University of Wisconsin-Milwaukee

December 2016

ABSTRACT

SYMMETRY AND RECONSTRUCTION OF PARTICLE STRUCTURE FROM RANDOM ANGLE DIFFRACTION PATTERNS

by

Sandi Wibowo

The University of Wisconsin-Milwaukee, 2016
Under the Supervision of Professor Dilano Kerzaman Saldin

The problem of determining the structure of a biomolecule, when all the evidence from experiment consists of individual diffraction patterns from random particle orientations, is the central theoretical problem with an XFEL. One of the methods proposed is a calculation over all measured diffraction patterns of the average angular correlations between pairs of points on the diffraction patterns. It is possible to construct from these a matrix B characterized by angular momentum quantum number l, and whose elements are characterized by radii q and q' of the resolution shells. If matrix B is considered as dot product of vectors, which magnetic quantum number m is the component, singular value of B reveals the number of magnetic quantum numbers in the spherical harmonics expansion. What is shown in this paper is dependency of magnetic quantum number on symmetry can be associated to lowest independent parameter to describe symmetry. At the very least this determines information about particle symmetry from experiment data, independent of

any assumed symmetry. An equally important point is that matrix B provides a means of reconstructing diffraction volume. This can be done by formulating intensity and matrix B as linear equation. Lastly, positivity constraint and optimization method is used to construct diffraction volume and phase is determined from phasing algorithm.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 XFEL

The world's first free electron laser of hard X-rays, which has been built in Stanford, is called the Linac Coherent Light Source (LCLS). It is a 4th generation X-ray source of ultra-short pulses of hard X-rays [2] built on the site of now abandoned instrumentation for particle physics. A capability of the LCLS is that it can create X-rays ten billion times brighter than those available before by any man-made source on earth, delivered at the rate of 120 pulses per second.

Since some crucial proteins such as membrane proteins are very difficult to crystallize, they may forever be outside the scope of traditional X-ray crystallography. The idea of using the XFEL for this purpose is to exploit its much greater brightness to obtain diffraction patterns of individual uncrystallized molecules. The fact that the brightness also means that the molecules are more likely to be destroyed by the incident beam is compensated by the fact that the peak brightness happens only over a period of the order of a femtosecond or so. The disintegration of the molecule takes at least 50 femtoseconds. Consequently, the x-ray scattering takes place while the molecule was in its original state. For the first time, this allows diffraction patterns (DPs) of undamaged samples to be

measured, essentially independent of dose. This principle is called "diffraction before destruction" and had been demonstrated experimentally [3].

However, there are definite differences with the practice of x-ray crystallography. A typical crystal studied by protein crystallography is perhaps 1 mm in linear dimensions. One would expect a typical protein to be perhaps 100 Å across. Thus one would therefore expect about 10^{15} molecules in a typical sample used in x-ray crystallography. One should therefore expect a typical diffraction pattern of a protein to be perhaps 10^{-5} weaker. Perhaps it is no wonder that the only reported experimental reconstructions by an XFEL is of the mimivirus which is perhaps a 10,000 Å cube. Thus one would perhaps expect an XFEL pattern of the mimivirus particle to be of similar intensity to a comparable crystal of mimivirus in a synchrotron.

For example, unlike the former case, the scattered intensities are not concentrated at Bragg spots, but are diffuse. This is due to the fact that the particles are randomly positioned and do not form crystals with perfect translational periodicity. This also rules out the use of orientation determination algorithms normally used in x-ray crystallography that are based on the idea of indexing.

In an XFEL, the particles are in random orientations and these do affect the intensities recorded in each diffraction pattern. This can be exploited to our benefit to help in generating a 3D diffraction volume by suitably orienting single-particle patterns. In other words the 3D diffraction volume can be thought of consisting of suitably oriented 2D diffraction patterns. The orientation must depend on the data in the diffraction patterns themselves. One idea is to use that fact similar patterns must be similarly oriented since if all molecules are identical and we concentrate on single particle pattern the only thing they can differ by is their orientations. This idea of similarity has been exploited in algorithms such as diffusion map. An alternative set of algorithms represent the diffraction volume in terms of a vector of each diffraction pattern as determined by components presenting the intensity at each pixel. If there are N such values per DP, due to the fact that each

diffraction has N pixels, these vectors will occupy an N dimensional space. In general this is too many dimensions to determine the orientations, which only need three parameters in SO(3). However dimensionality reduction techniques exist which find a suitably smooth 3D manifold from which the orientations may be read off, which have the added benefit of noise reduction An alternative method due to Elser [4] seems to find the orientations of patterns that give rise to a compact real-space object, via phasing. In fact most algorithms initially proposed for structure determination by an XFEL work with diffraction pattern from a single particle [4, 5] as determined by a hit-finder program [55]. A problem with hit-finder programs is that they tend to work with a small percentage of measured diffraction patterns. Indeed it has been estimated [7] that this percentage often amounts to 0.1 of 1 percent of the measured DPs, leading to the use of only 1 in a thousand of the amount of scarce proteins prepared, not to mention inefficiency of DP collection in an experiment. However, remember that an XFEL is only about 10^{10} brighter than a synchrotron source.

The use of a hit-finder program has another consequence, namely that the hit rate of single-particle diffraction patterns is only about 0.1 of 1 percent of all diffraction patterns measured [7] Nevertheless, such single particle methods have had some success even with experimental data, for example of the mimivirus [8] The theory developed in this dissertation is an attempt to develop theoretical methods that address this precise problem. While the vast majority of algorithms developed for this task proceed by finding the relative orientations of the molecules giving rise to the diffraction patterns it should be stressed that a relative angle is significant only if all molecules in the sample remain fixed relative to each other or else if one had only a single molecule contributing to a diffraction pattern. In reality if the molecules are presented to the X-ray beam in droplets over the course of several hours it is most likely different molecules will have changes to their orientations randomly as a consequence of molecular diffusion. The only thing that will stay constant is the molecular structure, and hence the molecular electron density. We will show that even in such cases we can deduce this structure from its angular correlations.

The integral of its angular correlations over all orientations will remain the same despite their possible different initial orientations, since they are an integral over all orientations the particle can have. Note this does not necessarily assume the particles are distributed evenly in angle. Even if particular orientations are favored, the sum of the orientations over all angles must be the same. Even if one had an ensemble of randomly oriented particles by the time one integrates over all orientations of the individual particles the contribution from each particle will be identical - it's just that the sum over different orientations is done in a different order - the sum over all possible orientations is identical. Consequently, if one sums over all diffraction patterns measured in an XFEL each molecule will have an identical contribution. One may call this the angular correlation function. Consequently, if one has a method for deducing a structure from its angular correlations it would work equally well from all randomly oriented particles, independent of their orientations of a particle in an individual diffraction pattern.

The problem is that the deduction of a particle's structure from its angular correlation function may be more difficult than its deduction from a single particle diffraction pattern, something that is well established nowadays by so-called iterative phasing programs. It's worth digressing a little to iterative phasing algorithms to understand this point. Due to the lack of phase information in measured intensities it is difficult to reconstruct a real-space density. However, there is no problem with going the other way. That is to say if one assumes an electron density one can always calculate a set of amplitudes by Fourier transforming the density and a set of intensities by taking the square moduli of the amplitudes. Once one has the intensity distribution one can calculate its spherical harmonic expansion and hence the Legendre transform of its pair (and ultimately) triple correlations. The idea then is to apply constraints in real and reciprocal space (the latter via the pair and triple correlations) to the same function or its Fourier transform to constrain to that function. In ordinary phasing algorithms the reciprocal space constraint is the intensity and the real space constraint is the support or approximate extent of the

electron density that may be known *a priori*. A normal phasing algorithm operates in a space described by a Cartesian coordinate system. In the XFEL problem, however the particles are presented to the beam in all possible orientations. Consequently, it is more appropriate to use polar coordinates. The reciprocal space constraints in this case is to the correlations. It is possible to write the correlations in terms of the spherical harmonic expansion coefficients of the diffraction volume. Consequently provided one works in a spherical harmonic system, there is not much difference from a phasing algorithm that constrains to an intensity in reciprocal space and one that constrains to a set of correlations. The real space constraint can remain the same support constraint as before. This is the essential idea behind the iterative phasing algorithm from the correlations proposed by Donatelli, Zwart, and Sethian in [5]. As an indication of its effectiveness, we show in the figures below, the electron density of the biomolecule directly from its PDB file in the left column and its reconstruction from the intensity correlations on the right. Obviously, the algorithm is really effective in performing the reconstruction.

Since the structure of a protein depends very much on whether or not it is in a hydrated environment, one method of delivery is of hydrated proteins to an XFEL within a solvent droplet of a few microns in size. Even if the beam vaporizes the protein and droplet, this does not matter in a diffract-and-destroy experiment [6] as it produces a diffraction pattern before then. If the background is constant, an argument due to Babinet may allow one to take account of the water scattering easily. Babinet [54] pointed out that of you add or subtract a constant density, it does not affect the sideways scattering, only the intensity normally located in the beam stop. Of course the assumption of constant water density is only valid at rather low resolutions, but so far XFEL experiments have only been done at resolutions worse than 100 Å . Intensities in a beam stop are normally estimated by allowing these intensities to float in an iterative phasing algorithm or by constraining these intensities by the known molecular weight of the protein. If one subtracts out exactly the value of the (constant) solvent density one would be assuming the

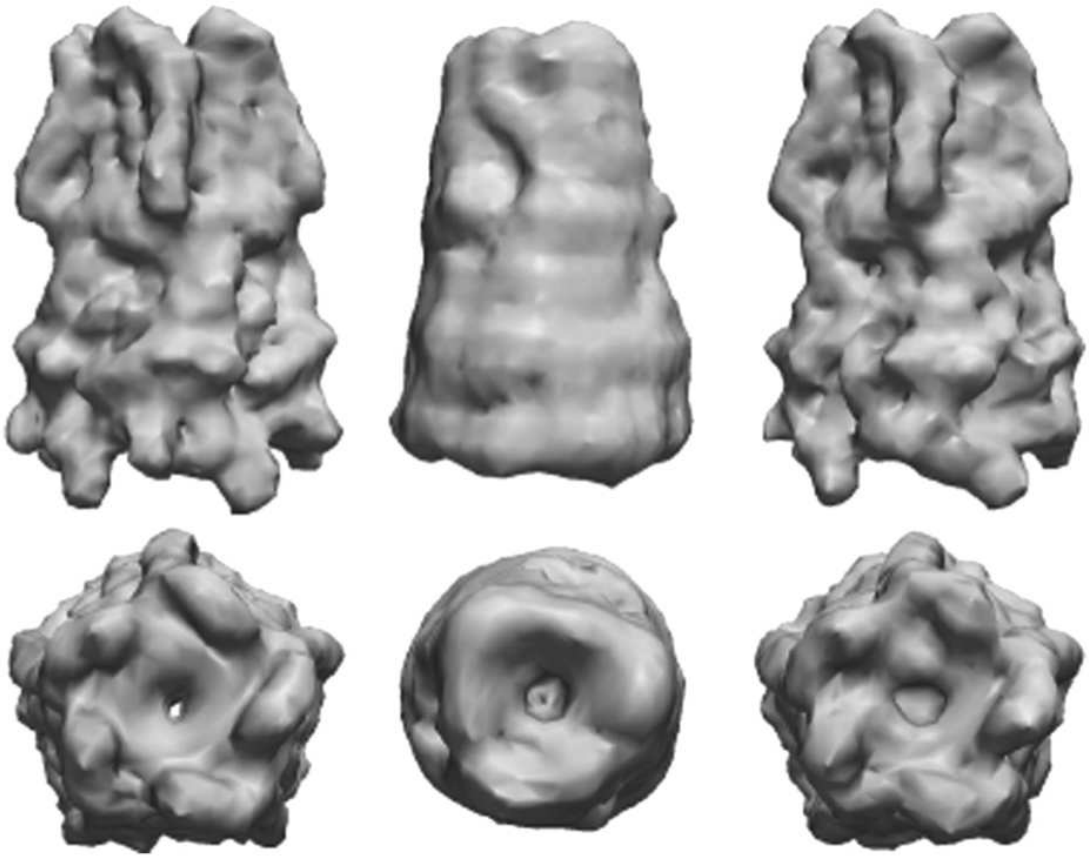


Figure 1.1: Protein electron desity reconstructed directy from the pair correlations by the M-TIP phasing algorithm derived by Donatell et al. [43]

scattering is by entities suspended in a vacuum but the electron densities of the entities had to be reduced by the solvent density. At least in the case of viruses, it has been possible to derive the structure from experimental data under this assumption. If it is possible to derive structure routinely from XFEL diffraction patterns which at the LCLS are measured at about 120 per second, then the possibility exists of measuring perhaps a million per experimental shift. The aim is to develop a method of extracting structural information from this data, to routinely solve the structures of biomolecules from the data. The XFEL unlocks the possibility of studying the structure of uncrystallized biomolecules [1]. Simulations show that molecules explode caused by intense brightness of the X-ray radiation after 50 femtoseconds beyond the initial incidence of the X-ray pulse [1]. However, meaningful diffraction patterns can be recorded before molecules explode

because the pulse produced by the XFEL is significantly shorter than the time needed for the molecular explosion.

A method was proposed originally by Zvi Kam [32] to obtain information about structure by correlating two points in each diffraction pattern and averaging over all diffraction patterns. This is completely logical in any circumstance where the orientations of the particles are unknown as the angular correlations do not depend on orientation in the same way that the usually measured intensities of scattering do not depend on particle position (and so the structure may be deduced from the intensities independent of particle position). Likewise, from the angular correlations, the structure can be deduced independent of particle orientation on any particular diffraction pattern as the structure deduction is from the sum over the data of all diffraction patterns, and the contribution of each particle is independent of its orientation in a particular X-ray pulse.

It is true that the number of particles whose diffraction patterns are sought will vary from shot to shot. However, this is of no relevance as one will form the pair correlations and triple correlations [12]

$$C_3(q, q', \Delta\phi) \;=\; \int I^2(q, \phi) I(q', \phi + \Delta\phi) d\phi \tag{1.1}$$

from exactly the same set of diffraction patterns. What is more, the pair and triple correlations will be identical in form independent of the number of particles. All that matters is that exactly the same set of diffraction patterns are used for the pair and triple correlations, which is easily enough arranged.

Crystallography is a method for determining structure of the molecular constituents of crystals [13]. X-rays hit large numbers of identical molecules arranged in a crystal and Bragg spots appear as a result of interference between the scattered X-rays. The intensities of the Bragg spots can be used to deduce the electron density of the molecule. The recovered density will be in a crystallized state whereas by using the XFEL, molecules

are shot in their noncrystalline state. By studying molecules in their noncrystalline states, one may gain further insight into how they function in nature.

Since individual biomolecules are studied in an XFEL, such objects have no translational symmetry and have no Bragg spots. What is more, as we have pointed out before, even their orientations are unknown. Despite this, we show that it is possible to deduce the structure from the collection of diffraction patterns measured in an XFEL. What is more, the angular correlations when integrated over all orientations are identical for all particles, Consequently, when a method is derived for reconstructing a structure from its correlations one should be able to deduce the structure of an individual molecule, even if a particular ensemble consists of many randomly oriented molecules [32]. We look in detail in this dissertation at the capabilities of the method of angular correlations. The reconstruction was done by simulating diffraction patterns from different random orientations of a virus that is known to have icosahedral symmetry [17] that is by simulating diffraction patterns known to be measurable in an XFEL. It should be stressed that all this method needs is a collection of diffraction patterns of random particle orientations. The flexibility of the method may be judged by the fact that it works just as well with diffraction patterns measured in the LCLS's Single Particle Initiative (SPI) [58] as with ensembles of randomly oriented particles that are probably inevitable with smaller molecules with a 1000 Å wide XFEL illumination area. it is assumed that the diffraction volume has icosahedral symmetry. Another important point is that by taking symmetry into account it will greatly reduce the number of independent parameters to construct the diffraction volume due to the fact that information on orientations of the particle is unknown.

Having information about the symmetry of particle is valuable. When transformed by a Legendre function P_l, the pair correlations. C_2 gives rise to a quantity B_l that depends on the angular momentum quantum number l [8]. It is possible to use the information contained in B_l (and T_l a similar transform of C_3) to deduce the magnitudes and signs of spherical harmonic coefficients of the diffraction volume characterized by l but not by

the magnetic quantum number m [17]. Until recently, information about m has to be deduced by the known symmetry of the particle As a result of this work at least the number of m values may be found from the B_l's (derivable from experiment by singular value decomposition). Also the recent, as yet unpublished work of Donatelli, Saldin, and Zwart suggests a method of finding the full $I_{lm}(q)$ coefficients from the correlations.

While on the subject of the Single Particle Initiative [10], this is an attempt at the LCLS to collect XFEL data from single molecules, but of all possible orientations to within a Shannon angle (The Shannon angle is defined as roughly the width of a single feature in a diffraction pattern). In order to facilitate experiments that hit single particles, initial experiments have been on large bioparticles such as viruses. Consequently, we applied our method to experiments with the so-called rice dwarf virus (RDV) [14] conducted at the LCLS in August 2015. The results are shown here. This shows a computer reconstructed image with a computational slice made to indicate whether or not there is genetic material on the inside. Our image correctly showed the genetic material inside unlike a structure reconstructed (also shown) from the data in the Protein Data Bank where the internal genetic material was removed. Luckily, for a symmetry based method, for the two main categories of regular virus the icosahedral [17] and helical [39] a knowledge the symmetry allows a complete solution. The symmetry is an assumption taken in order to fill missing steps of the reconstruction. We have already seen that a knowledge of a particles symmetry is of great help in determining some of the crucial quantum numbers in the angular momentum description. Ideally one would like to determine these symmetry parameters from the experimental data rather that by assumption. We how in this dissertation that this is indeed possible by a singular value decomposition of quantities $B_l(q, q')$ deducible from the angular pair correlations.

From a study of angular correlations, virus structure can be reconstructed by constraining intensity to be always positive. Under icosahedral symmetry, only the signs of the spherical harmonics expansion of the diffraction volume are not unique from the B_ls

and a positivity constraint is suitable to resolve the signs. It will be shown in this dissertation that a positivity constraint can be used to determine the intensity from the matrix $B_l(q, q')$ by an optimization method thus enhancing the constraint to be used not just in icosahedral symmetry but also for different types of symmetry. For example, Caspar and Klug [31] once said all regular viruses fall into the symmetry classes of icosahedral or helical so there are in any case not many symmetries to be tried. It would be helpful if these symmetries are deduced directly from the data as we show in this dissertation, and not left to trial and error.

Although the method of correlations is useful for getting more information from situations where one only has incoherent data, it is also useful for coherent XFEL radiation for a disordered array. What we mean is that in the presence of multiple particles one needs

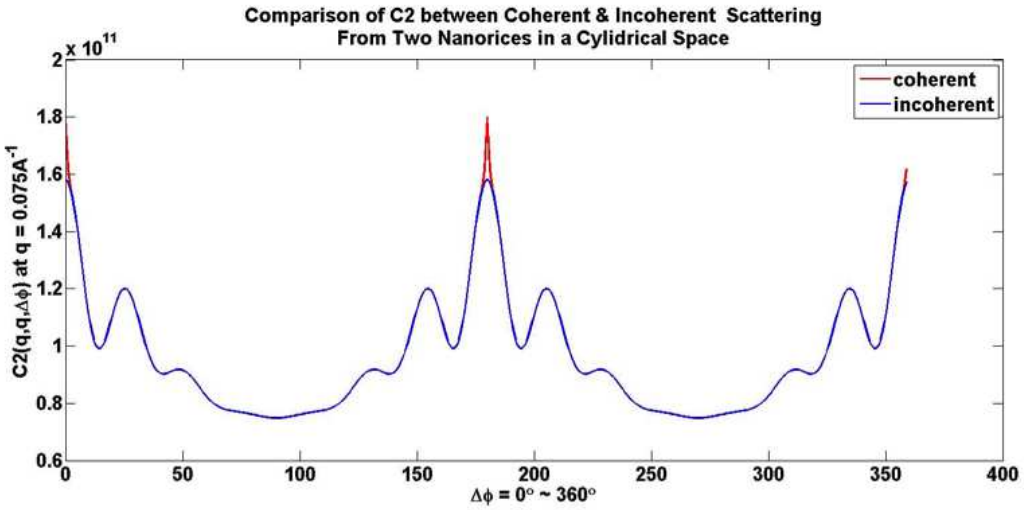


Figure 1.2: Coherent peaks in (in red) in the correlations from incoherent diffraction patterns from the contributions of two independently randomly oriented nanoparticles, because the disorder gives rise to a kind of incoherence (except for narrow regions of reciprocal space that can easily be ignored)

to look at correlations between particles as a result mutual interference. Quite simply, in the presence of two particles in a source of coherent radiation such an XFEL one would expect the total intensity to be $|\sum_{j=1} F_j \exp(i\vec{q} \cdot \vec{r_j})|^2$ where F_j is the structure factor

of the jth molecule. The exponentials give rise to a factor of $\exp(i\vec{q}\cdot(\vec{r}_j - \vec{r}_k))$ which results in random phases if the particle positions are random except that as $q \to 0$, when all phase factors become zero and are thus not random. Thus over most of its range one would expect interference fringes perpendicular to $\vec{r}_j - \vec{r}_k$ if the radiation is coherent. Fortunately, for different atom pairs, these fringes are random in orientation and spacing which makes the sum of cross terms amongst different particles tend towards zero, making the radiation effectively incoherent, over most of the q range as pointed out before. It would be of interest though if the interference fringes exist. This is precisely what is observed in Figure 1.3.

While on individual diffraction patterns, fringes due to interference between the two particles are visible, the randomness of particle positions means that on different diffraction patterns these fringes will be in random directions and of random spacings. Consequently. when one adds contributions from different particles of random positions, the fringes essentially average out and it is if the sum is incoherent. That is, it is as if one were summing patterns like that in the bottom left above. Due to the randomness of the phases one may ignore the second summation over most of the q range and therefore over most of the range one obtains what one would be equivalent of the incoherent sum $\sum_j I_j$ where I_j is the intensity scattering contribution from particle j. Thus the total intensity reduces to a sum of intensity contributions from particle j, as if the scattering was not coherent [60]. The only exception occurs near $\Delta\phi = 0$ (see Fig. 1.2, and equivalently $\Delta\phi = \pi$ due to Friedel symmetry, and $\Delta\phi = 2\phi$ (same as $\Delta\phi = 0$). These peaks are due to the fact that near $\Delta\phi = 0$, all scattering phases become equal (and equal to zero). Thus the assumption of random phases is no longer valid. However, the width of such a peak is of the order of $2\pi/L$ where L is the width of the coherent radiation (about $L = 1000$ Å at the LCLS). Thus $2\pi/L$ is usually much smaller than the width of a Shannon pixel π/D where D is of the order of 50 Å . Thus, in calculating B_l or T_l by integration under the curves of C_2 and C_3, respectively, can ignore the sharp high coherent peaks and still

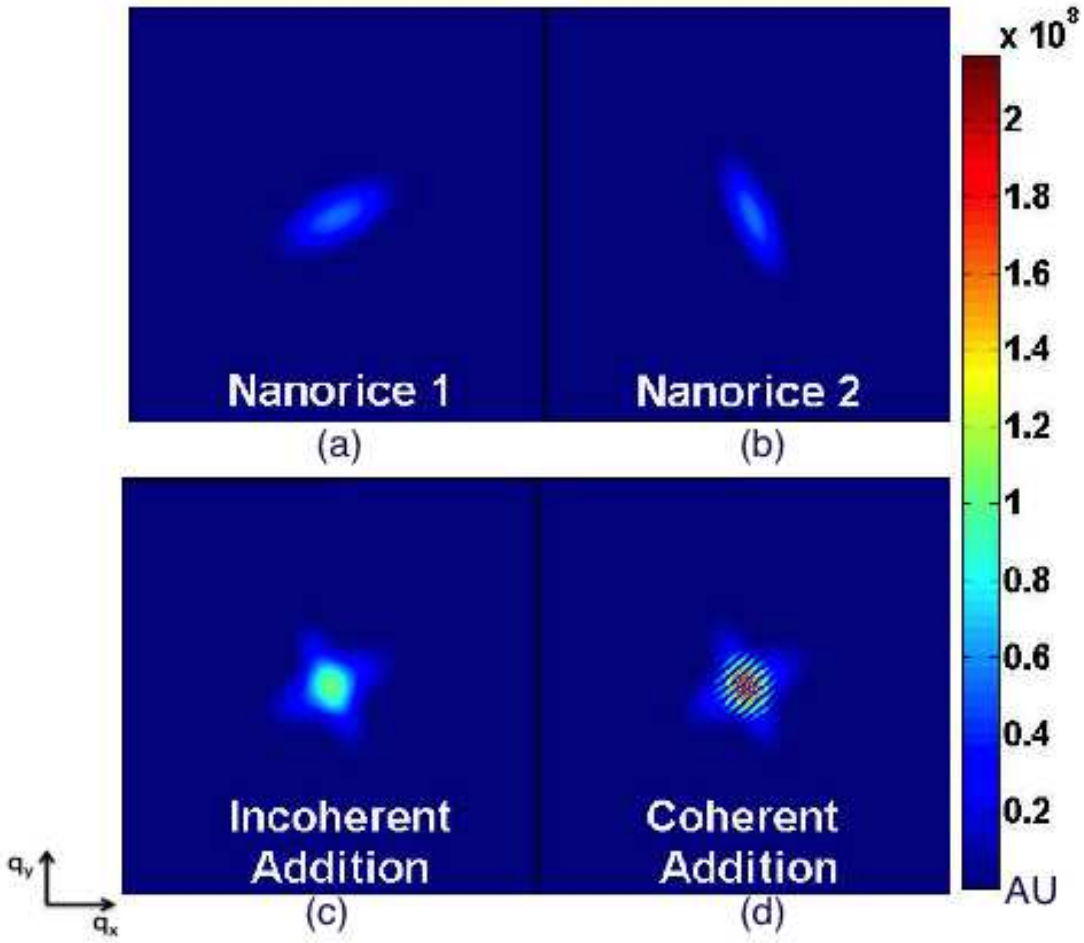


Figure 1.3: (a) and (b) are single particle diffraction pattern in different orientation, (c) is incoherent diffraction pattern and (d) is coherent diffraction pattern. If the radiation is coherent, one will see interference fringes, which will average out of there are many particles of random position.

get essentially the same result. Thus the conclusion is for the present application of the reconstruction of the electron density of a biomolecule or virus from XFEL coherent radiation, these narrow peaks can be neglected, and the previous theory [33] that applies also the single particle experiments like in the Single Particle Initiative [58] is applicable.

The method of angular correlations is also of great help with helical viruses [39]. In the past it has been attempted to study these entities by aligning them as in a fiber

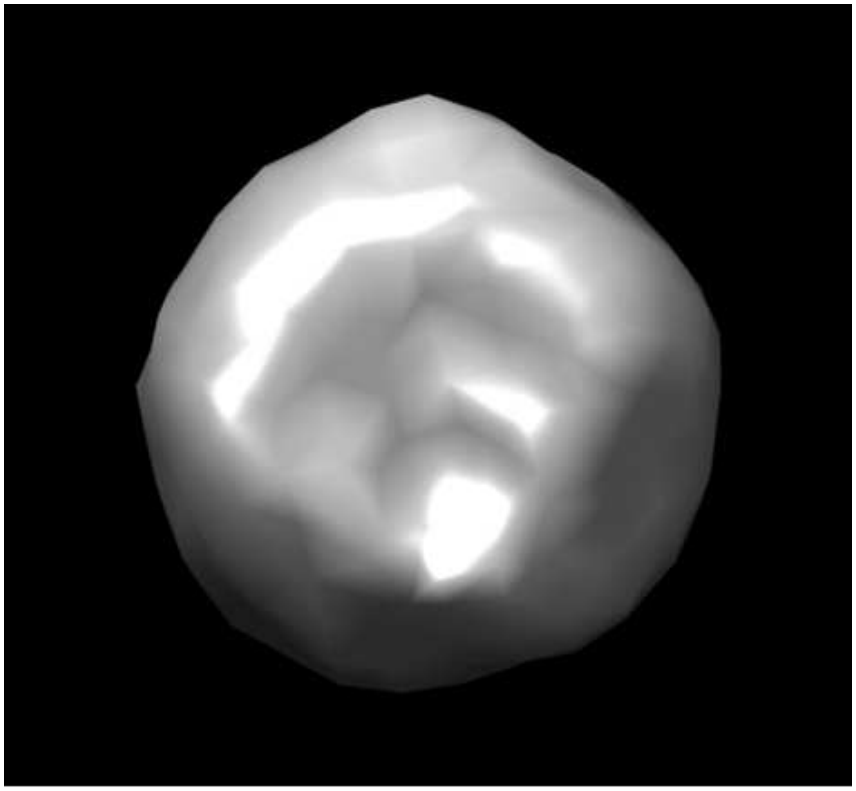


Figure 1.4: The rice dwarf virus (RDV) reconstructed from experimental data from the Single Particle Initiative measured in August 2015. Note the apparent existence of internal genetic material, as the viruses in this experiment did not have the internal genetic material removed

by physical means such was powerful electric fields. This has always run up against the obstacle of the entropic tendency to misalign.

Since the orientation of the reconstructed image may be chosen arbitrarily this allows an opportunity to use of the correlation method to align helical viruses computationally. It is usually assumed that the diffraction volume may be characterized by a magnetic quantum number $m = 0$ if the z-axis can be taken along the helix. It turns out that even if the helices are randomly oriented in practice merely choosing $m = 0$ for the spherical harmonic components of the pair correlations B_l, computationally aligns the helical viruses and allows an estimate of the values of the spherical harmonic expansion coefficients of the diffraction volume [39]. Even if this is regarded as an approximation, the perturbation method we have developed for time-resolved structure [35] is capable of

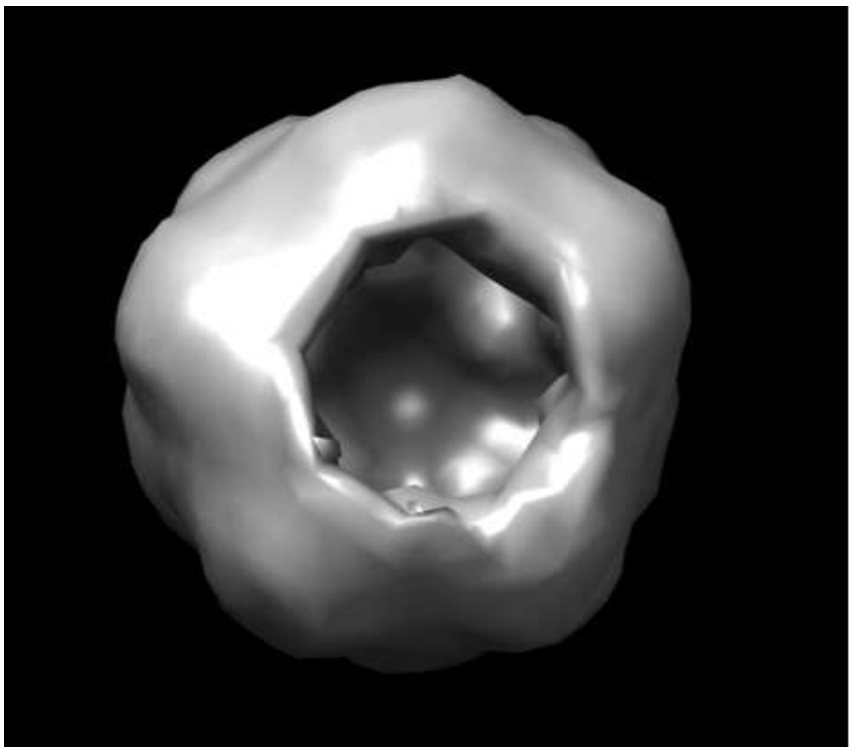


Figure 1.5: Similar image of the satellite tobacco necrosis virus whose structure in deposited in the protein data bank. This has had its internal genetic material removed, as revealed by the reconstructed image

refining the values.

A real advantage of our method over all others that have been proposed for this problem is that it reconstructs the image from its correlations, Since the angular correlations of randomly oriented particles are identical, one can reconstruct an image of a single particle from an experiment consisting of multiple randomly oriented particles. Since the angular correlations are the same, independent of particle orientations, a corollary is that it may be reconstructed in any orientation. In general, an orientation is chosen to be consistent with the representation of the particle. An image of a single particle of nanorice reconstructed from diffraction patterns of two randomly oriented particles is shown next. In the case of a helical virus or a particle of nanorice, the diffraction volume is assumed to be azimuthally symmetric and m=0 is the only permitted component of the magnetic quantum number. (It should be emphasized that this is only possible because of the property

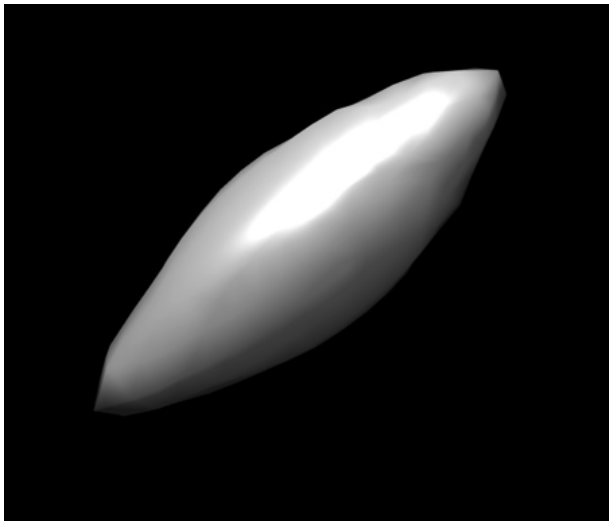


Figure 1.6: Single particle of nanorice reconstructed from diffraction patterns of two independently randomly oriented particles.

of angular correlations as being the same independent of the particle orientation.)

With a focal spot of 1000 Å, it is quite hard to focus on a single particle, and most diffraction patterns of proteins will probably be from multiple particles. It is true that one may remove diffraction patterns from multiple particles by so-called hit finder methods. But this is only at the expense of hit rate, as we have commented earlier

It should be mentioned that, as currently formulated, the quantities B_l and T_l derived from C_2 and C_3, respectively, depend only on the azimuthal quantum number l. whereas the general the spherical harmonic expansion coefficients of the diffraction volume are characterized by both l and the magnetic quantum number m. Consequently, it was proposed for both icosahedral and helical viruses that one uses the known symmetry properties for deducing the value of m [32].

Ideally of course one may need to apply this method to completely non-symmetric particles. It has recently been shown to be possible to obtain spherical harmonic coefficients $I_{lm}(q)$ characterized by particular values of m by using the fact the so-called 3-point triple correlations. One first calculates the $I_m(q)$ coefficients of a circular harmonic expansion of the projections the structure using the method of Kurta et al. [61] and Pedrini et

al. [62]. Of course as one goes to lower X-ray energies one can exploit the increasingly curved nature of the Ewald sphere to get information on the $I_{lm}(q)$ coefficients of the 3D diffraction volume by an experiment like one on a black-lipid membrane. This will be no problem for membrane proteins which like to live within a membrane anyway. Since one of the stated aims of XFEL work is to determine the structure of hard-to-crystallize membrane proteins this a fulfillment of one of the original aims of the construction of a nearly billion dollar XFEL.

We should also mention here other advantages of an angular momentum method particularly for icosahedral structures. Of the angular momenta l, while $l = 0$ obviously has icosahedral symmetry, the next higher value of l consistent with this symmetry is $l = 6$. Consequently, if B_l values are found from experimental data, the lower l values should be dominated by $l = 0$ and $l = 6$. Thus even without an assumption of icosahedral symmetry one can get some indication of such symmetry form the experimental data even without a reconstruction of the particles image in real space. An example of such a calculation is shown below.

Another advantage concerns the values of the intensity in the beam stop. There are less and less angular momenta as the scattering angle is reduced, In fact it can be shown that the maximum value of l associated the outer edge of the beam stop is about 5. Since the maximum angular momentum associated with a given radius on a diffraction pattern is proportional to the radius and given that fact that the next lower angular momentum value consistent with icosahedral symmetry is $l = 0$ one can estimate the intensity inside the beam stop if one has an analytic expression of the intensity that is angularly symmetric. In fact the known analytic form of the intensities from a uniform sphere of scattering matter is angularly symmetric, and it can be assumed to be the analytic extension of the computed intensities at higher scattering angle. This can be used to extend some of the intensities into the beam stop provided one ensures that the radial part of the data are continuous between the outer computational part and the inner

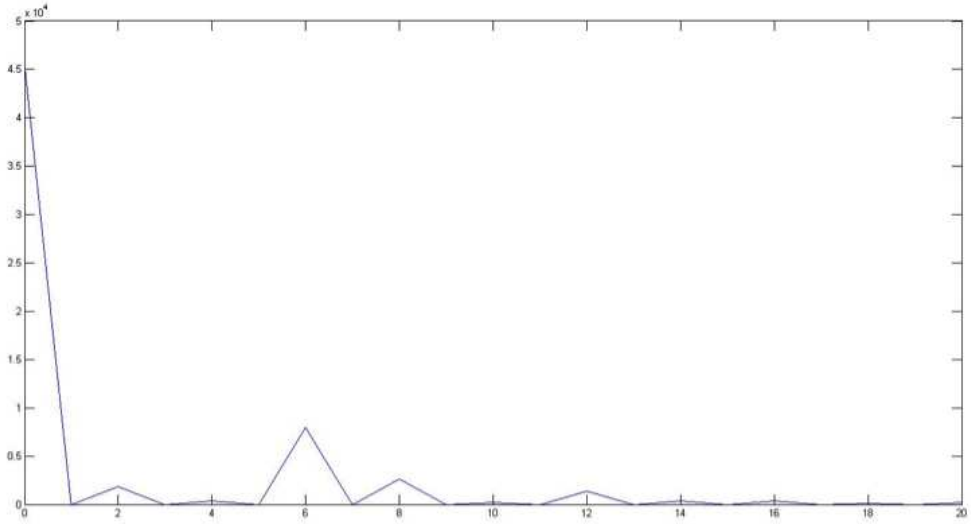


Figure 1.7: Calculation of the values of B_l from experimental diffraction data from the rice dwarf virus without any symmetry assumption. This is dominated by $l = 0$ and $l = 6$, a signature of icosahedral symmetry.

analytic part. Indeed flipping-based phasing algorithms [29, 30] are often very sensitive to the extent of the beam stop, and the extension of the data by this means is often of great help with a phasing algorithm.

Chapter 2

Theoretical Foundation

2.1 X-ray Diffraction

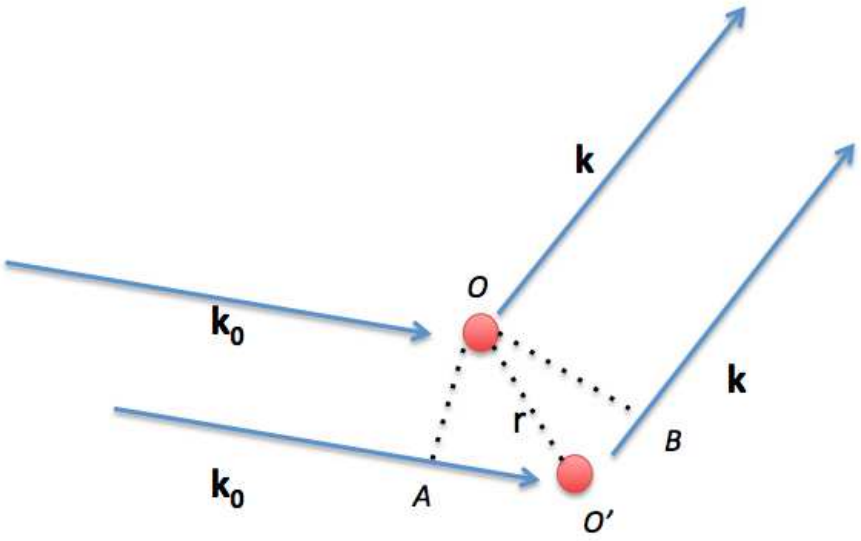


Figure 2.1: Diagram of X-ray diffraction

The diagram in Figure 2.1 shows the relation between the incoming waves, the scattered waves and the phase difference. The incoming wave that has wave vector $\vec{k_0}$ hits two electrons and they are scattered with the direction of $\vec{k}$. The scattered waves are parallel

each other under approximation the observer is very far.

Because the scattered waves are scattered at different positions, the scattered waves will have a phase difference. Another way to see this, the phase difference arises because each scattered waves travel a different length. From figure 1, the bottom wave travels longer than the top wave so that there is a difference in path length. From figure 2.1, the difference in path length is

$$\text{Path difference} = \text{AO'} + \text{O'B}. \tag{2.1}$$

AO' is projection of $\vec{r}$ along $\vec{k_0}$ and has length $\vec{r} \cdot \vec{k_0}$. On the other hand, O'B is negative projection of $\vec{r}$ along $\vec{k}$ and has length of $-\vec{r} \cdot \vec{k}$. The total path difference is $\vec{r} \cdot (\vec{k_0} - \vec{k})$ or $\vec{r} \cdot \vec{q}$ where $\vec{q}$ is $(\vec{k_0} - \vec{k})$. The total phase difference become $\exp(2\pi\vec{r} \cdot \vec{q})$.

The diffraction multiplies the amplitude of the scattered wave by a phase factor $\exp(2\pi\vec{r} \cdot \vec{q})$. If there are many electrons with density $\rho(\vec{r})$ then the effect at particular point $\vec{q}$ will sum to

$$A(\vec{q}) = \int \rho(\vec{r}) \exp(2\pi i \vec{q} \cdot \vec{r}) d\vec{r}. \tag{2.2}$$

So the structure factor appears as a Fourier transform of the electron density. The diffraction experiments only measure the square of absolute value of $A(\vec{q})$, which shows up as the intensity corresponding to $\vec{q}$. Mathematically, the intensity can be written as

$$I(\vec{q}) = |A(\vec{q})|^2. \tag{2.3}$$

There is a more convenient way to calculate a structure factor of the molecule rather than perform Fourier transform of its full electron density. The structure of the molecule can be decomposed into its individual atoms. As already known, there are many the same type of atoms inside the molecule but they differ in positions only. By knowing the

Fourier transform of a single type of atom, it leads us to have easier computation because the total structure factor is a sum over all contribution of the Fourier transform of atoms in all position. Thus, calculating the Fourier transform of a single atom enables one to perform easier simulations to calculate structure factors.

The Fourier transform of a single atom is called atomic form factor. Based on a work done by Don Cromer and Mann [48], the Hartree-Fock approximation can be used to obtain empirical parameters to approximate atomic form factors. The way they determined the parameters was by fitting 9 parameters in a Gaussian's series to a normalized scattering curves. Currently, those parameters are readily available from the international table of crystallography [49] and the Gaussian function is shown in equation 2.4.

atom	a_1	a_2	a_3	a_4	b_1	b_2	b_3	b_4	c
C	2.31	1.02	1.589	0.865	20.84	10.21	0.569	51.65	0.216
N	12.213	3.132	2.013	1.166	0.006	9.893	28.997	0.583	-11.529
O	3.049	2.287	1.546	0.867	13.277	5.701	0.324	32.909	0.251
S	7.070	5.340	2.236	1.512	1.366	19.828	0.092	55.228	-0.159

Table 2.1: Table of Cromer-Mann coefficients

The parameters for different type of atoms are listed in table 2.1. There are 9 parameters for each atom and the table shows only entries for carbon, oxygen, nitrogen, and sulfur. After knowing all 9 parameters, the atomic form factor can be calculated using Gaussian function:

$$f(\sin(\theta)/\lambda) = \sum_{i=1}^{4} a_i \exp(-b_i(\sin(\theta)/\lambda)^2) + c. \tag{2.4}$$

A plot of the atomic form for carbon and oxygen is shown in figure 2.2. It is shown in the plot that the value of the atomic form factor goes to their atomic number when $\sin(\theta)/\lambda$ close to zero.

The structure factor can be calculated in a simpler way if the approximation of atomic form factor is used. Because the atomic form factor is calculated once, the calculation of

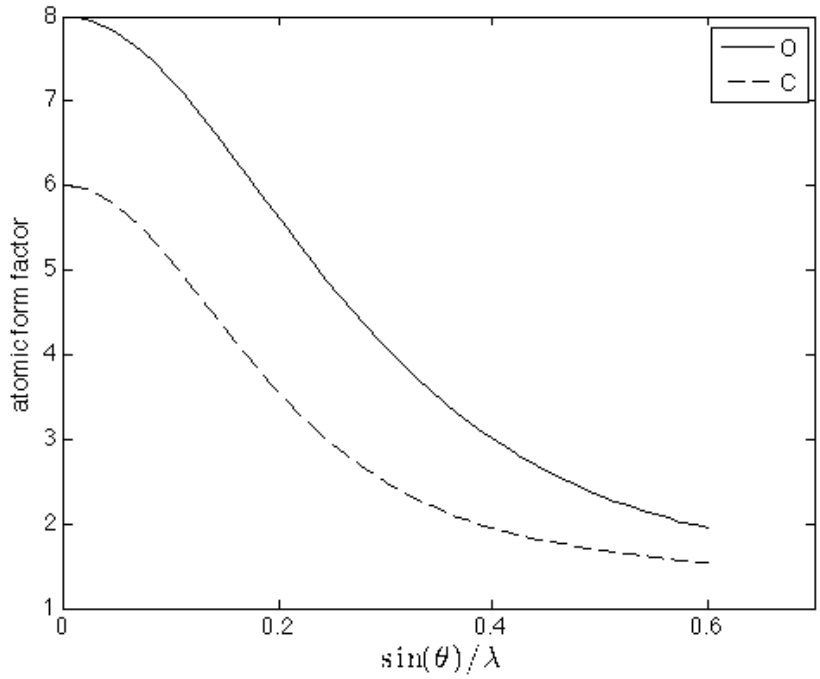


Figure 2.2: Plot of atomic form vector for carbon and oxygen

the structure factor is done faster for all atoms. Finally, the expression for the structure factor in terms of the atomic form factors is described as

$$A(\vec{q}) = \sum_i f_i(q) \exp(2\pi \vec{q} \cdot \vec{r_i}). \tag{2.5}$$

The equation 2.5 will be used to simulate the structure factors for a molecule. As long as a molecule is listed as a collection of atoms in different positions, then equation 2.5 can be used to simulate the structure factor. Some structures of molecules have been solved using methods of crystallography and their structures are available in the protein data bank (pdb). The pdb file describes a molecule as a list of atom type as well as their positions. Therefore, one can simulate a structure factor by using equation 2.5 where the entry is from the pdb file.

Figure 2.3 is a snapshot of a part of the pdb file. In order to read the information from pdb file, one requires to understande thoroughly the format and the convention of

```
ATOM      1  N   THR A  12      67.946  33.337  16.826  1.00102.16           N
ATOM      2  CA  THR A  12      66.915  34.315  17.328  1.00102.16           C
ATOM      3  C   THR A  12      66.570  35.274  16.204  1.00102.16           C
ATOM      4  O   THR A  12      66.803  36.487  16.305  1.00102.16           O
ATOM      5  CB  THR A  12      67.518  35.046  18.502  1.00102.16           C
ATOM      6  OG1 THR A  12      66.587  35.986  19.020  1.00102.16           O
ATOM      7  CG2 THR A  12      68.790  35.815  18.139  1.00102.16           C
ATOM      8  N   MET A  13      66.002  34.684  15.140  1.00 89.79           N
ATOM      9  CA  MET A  13      65.641  35.451  13.882  1.00 89.79           C
ATOM     10  C   MET A  13      64.509  36.438  14.049  1.00 89.79           C
ATOM     11  O   MET A  13      64.222  37.242  13.151  1.00 89.79           O
ATOM     12  CB  MET A  13      65.134  34.499  12.528  1.00102.16           C
ATOM     13  CG  MET A  13      64.734  35.183  11.118  1.00102.16           C
ATOM     14  SD  MET A  13      64.160  34.185   9.729  1.00102.16           S
ATOM     15  CE  MET A  13      63.807  35.229   8.332  1.00102.16           C
ATOM     16  N   ARG A  14      63.810  36.459  15.148  1.00 50.58           N
ATOM     17  CA  ARG A  14      62.716  37.444  15.257  1.00 50.58           C
ATOM     18  C   ARG A  14      63.265  38.810  15.654  1.00 50.58           C
ATOM     19  O   ARG A  14      62.901  39.839  15.065  1.00 50.58           O
ATOM     20  CB  ARG A  14      61.689  37.045  16.311  1.00102.16           C
```

Figure 2.3: Example of data from protein data bank in pdb format

the file. First, the pdb file has row entries where each row is a single atom in particular position together with additional information. It consists of multiple columns where each column has particular information. In total, there are 27 columns and all data is in a text file in ASCII format.

For the purpose of simulating the structure factors, only atom types and their positions are needed. Thus, there are four pieces of information needed, namely atom type, position-x, position-y, and position-z. The atom type is shown between columns 13 to 16. The position-x is shown between columns 31-38. The position-x is shown between columns 39-46. The position-x is shown between columns 47-54. With the information above, the structure factor can be simulated using equation 2.5 with the source of a pdb file. Full explanation about the format of pdb file is given in appendix C

2.2 Angular Correlation

A single particle diffraction experiment is an experiment that diffracts individual biomolecules using high intensity X-rays without crystallization. Figure 2.4 shows the schematic design of the experiment. The incoming X-ray produced in LCLS has high enough intensity so that detector can capture the scattered waves. The injector is capable of streaming the

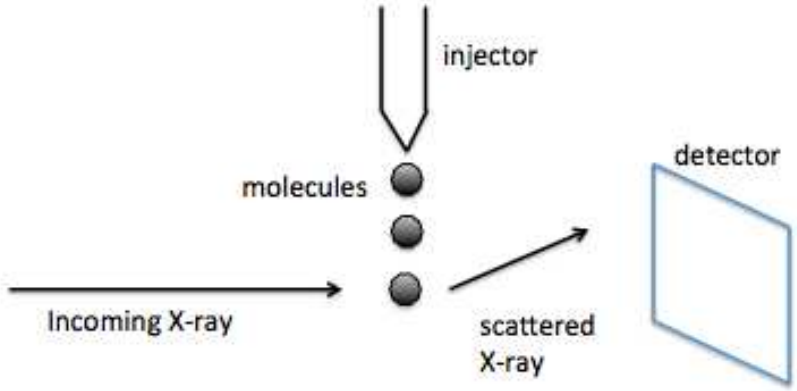


Figure 2.4: Diagram of single particle diffraction experiment

molecules in a tiny diameter so that there is chance an X-ray will hit a single molecule.

The information obtainable from this setup is the diffraction patterns of the molecules. However, there is missing information from the setup, namely the information about the orientations of the molecules. Each diffraction pattern recorded by the detector is very noisy therefore cannot be used for information about the orientation of the molecules. It is important to note that the detector is able to record many millions of diffraction patterns. Although the information about the orientations is lost, it is still possible to get the information about the structure of the molecule by averaging many diffraction patterns. The next section explains the theory to reconstruct the structure of the molecules by averaging many random orientations of the diffraction patterns.

Figure 2.5 illustrates typical outputs of a diffract and destroy experiment. The output consists of a collection of the diffraction patterns in random orientations. In order to remove the angular dependence, we need to take average over all diffraction patterns. Because the structure of molecule cannot be obtained by only taking average of a point in the diffraction patterns as all point average to the same for random molecules orientations, two point averaging is done to obtain more information about the structure of molecules. The final goal is to derive an orientation-independent quantity, which has information

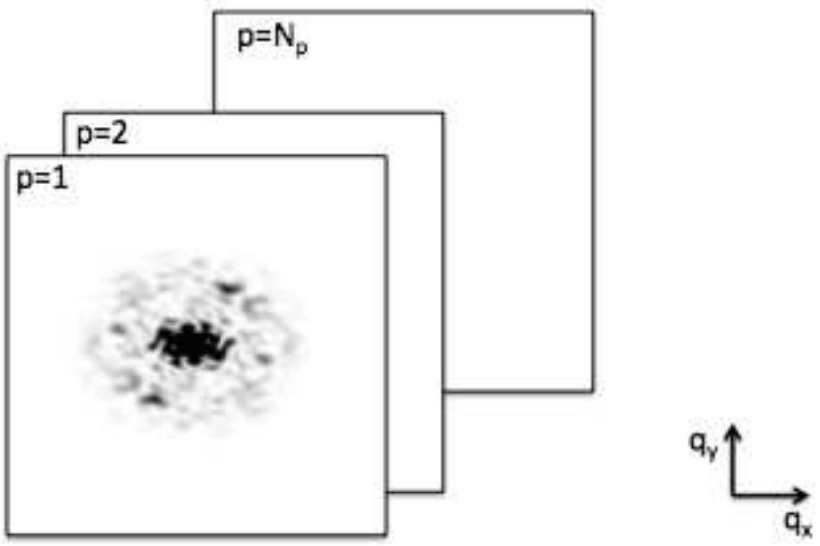


Figure 2.5: Collection of random angle diffraction patterns

about the structure, by correlating two points in the diffraction patterns and summing over all diffraction patterns.

Before going into the derivation of correlations, it is important to derive a relation between the intensity and the diffraction patterns. Figure 2.6 is a section through the Ewald sphere and a single diffraction pattern samples 3D reciprocal space in Ewald sphere. Consequently, one can derive the relation between the polar angle θ and the distance q, namely

$$\theta(q) = \frac{\pi}{2} - \sin^{-1}(\frac{q}{2\kappa}) \tag{2.6}$$

as illustrated in figure 2.6.

The curvature of Ewald sphere for arbitrary X-ray wave number κ is taken into account correctly by expressing θ in terms of q and κ. By substituting θ in equation 2.6, any point in a diffraction pattern can be specified by its q and ϕ as illustrated in figure 2.7.

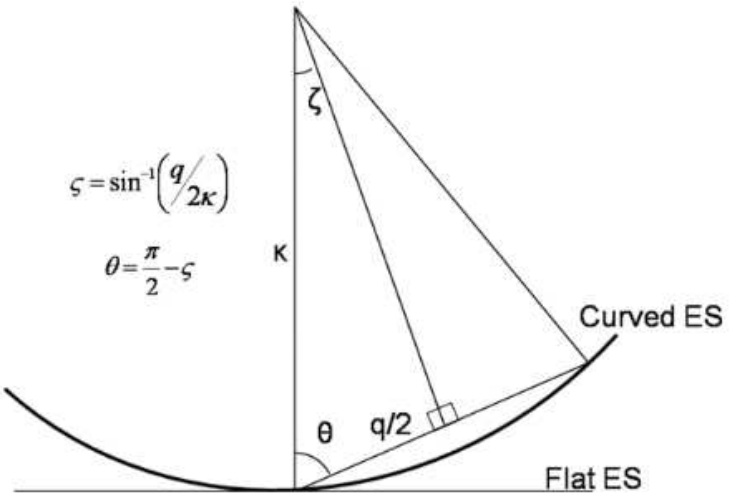


Figure 2.6: Relation between reciprocal radial distance q and angle θ in an Ewald sphere [33]

Another step is by taking Z-axis as the direction antiparallel to the incident wave; then the measured intensity in a diffraction pattern can be expressed as

$$I_Z(q,\phi) = \sum_{lm} I_{lm} Y_{lm}(\theta(q),\phi). \tag{2.7}$$

Figure 2.5 illustrates that there are many diffraction patterns and index p corresponds to the diffraction patterns with different molecular orientations. The orientation can be seen as a rotation of frame of reference because a rotation of the molecule is equivalent to an inverse rotation of its frame of reference. Mathematically, the particular orientation can be expressed by applying rotation operator to its original basis function. Specifically, the rotation operator is matrix D_{lm} because we chose spherical harmonics as basis function of intensity. Consequently, the new diffraction pattern in rotated frame of reference is

$$I^{(p)}(q,\phi) = \sum_{lmm'} I_{lm}(q) D^{(p)}_{lmm'}(\alpha,\beta,\gamma) Y_{lm'}(\theta(q),\phi) \tag{2.8}$$

where p is the index of the diffraction patterns as shown in figure 2.5 and (α,β,γ) are Euler angles.

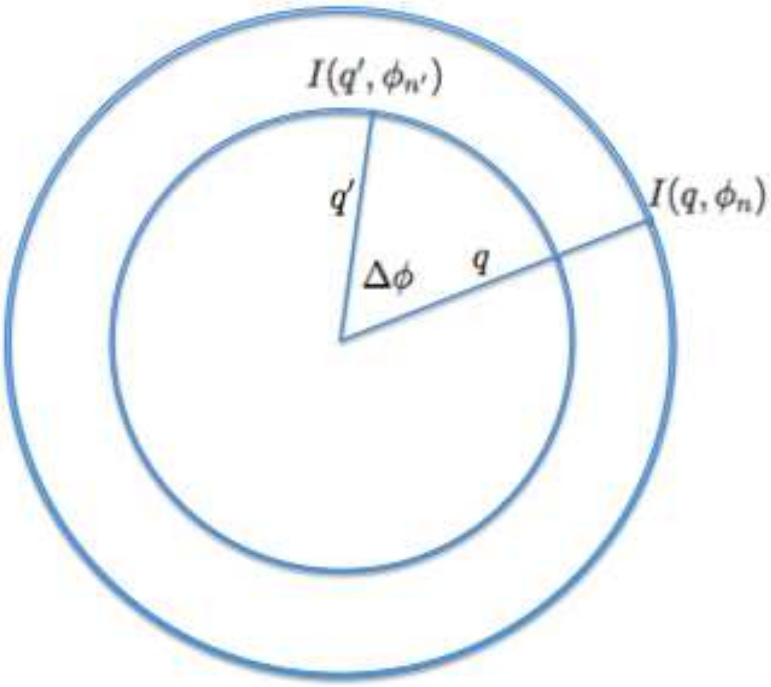


Figure 2.7: Two-point-correlation in a diffraction pattern

The first step in using this method is to calculate angular cross correlations on each diffraction pattern in polar coordinates. Polar coordinates are natural for this problem since the particles differ mainly in their orientations (They may also differ in position, but this does not affect the diffraction pattern intensities that are insensitive to the particle phases).

As illustrated in figure 2.7, we can pick any two points in the polar diffraction pattern by specifying the coordinate q and angle ϕ. The next step is to correlate every point in rings q and q' by keeping the same angular distance ϕ and ϕ'. Angular pair correlations are defined by

$$C_2(q, q', \phi, \phi') = \frac{1}{N_p} \sum_p I^{(p)}(q, \phi) I^{(p)}(q', \phi') \tag{2.9}$$

where p is the index of the diffraction patterns and N_p is the total number of diffraction pattterns as illustrated in figure 2.5.

Equation 2.9 can be expressed in terms of a summation of points inthe diffraction patterns rotated by matrix $D^{(p)}_{lm}$. By substituting equation 2.8 into equation 2.9, the C_2 become

$$C_2(q,q',\phi,\phi') = \frac{1}{N_p}\sum_{p}\sum_{lmm'}\sum_{l'm''m'''} I^*_{lm}(q)D^{(p)*}_{lmm'}Y^*_{lm'}(\theta(q),\phi) \times I_{l'm''}(q)D^{(p)}_{l'm''m'''}Y_{l'm'''}(\theta'(q'),\phi') \tag{2.10}$$

The Wigner D-matrices are representation of the full rotation group. A set of the Euler angles specify the rotation of matrix D. Due to the randomness of the orientations of the diffraction patterns, the larger the number of diffraction patterns the most likely the angles will occupy the entire space. Under assumption that the set of random angles will converge into all uniform rotational angles then equation 2.10 can be simplified. The relation that is used to simplify the equation is called the great orthogonality theorem, which is mathematically expressed as

$$\frac{1}{N}\sum_{(p)} D^{(p)*}_{lmm'}D^{(p)}_{l'm''m'''} = \frac{1}{2l+1}\delta_{ll'}\delta_{mm''}\delta_{m'm'''}. \tag{2.11}$$

By summing first over p in equation 2.10, making use of the great orthogonality relation in equation 2.11, and then summing over l', m'', and m''' will transform equation 2.10 into

$$C_2(q,q',\phi,\phi') = \sum_{l} F_l(q,q';\phi,\phi')B_l(q,q') \tag{2.12}$$

where

$$F_l(q,q';\phi\phi') = \frac{1}{2l+1}\sum_m Y^*_{lm}(\theta(q),\phi)Y_{lm}(\theta'(q'),\phi') \quad (2.13)$$

$$= \frac{1}{4}P_l[\cos\theta(q)\cos\theta(q') + \sin\theta(q)\sin\theta(q')\cos(\phi-\phi')] \quad (2.14)$$

where P_l is a Legendre polynomial of order l, and

$$B_l(q,q') = \sum_m I_{lm}(q)I^*_{lm}(q'). \quad (2.15)$$

The left hand side of equation 2.12 is obtainable from experiment. The first term of right hand side of equation 2.12 can be calculated mathematically. Consequently, the quantity B_l can be obtained from experiment; it can be used to get the information about the structure of the molecule.

The calculation to extract B_l from equation 2.12 is matrix inversion. For each pair q and q', equation 2.12 may be written as the matrix equation

$$C_{2(\phi\phi')} = \sum_l F_{\phi\phi',l}B_l. \quad (2.16)$$

All elements of matrix F are real numbers. Thus, the above equation can be inverted to get real coefficients B_l where

$$B_l = \sum_{\phi\phi'} {F^{-1}}_{l,\phi\phi'}C_{2(\phi\phi')}. \quad (2.17)$$

The above equation can be used to calculate $B_l(q,q')$ after C_2 is obtained. The information about the structure of the molecules is contained in $B_l(q,q')$ because $B_l(q,q')$ containis information about $I_{lm}(q)$ where $I_{lm}(q)$ are spherical harmonic expansion coefficients of a diffraction volume. Thus, the information about the structure of molecules can be ob-

tained by calculating $B_l(q, q')$ from a set of randomly-oriented diffraction patterns.

The spherical harmonics are used to expand the diffraction volume because it can construct any function in a 2D surface. In the 3D case, a molecule is free to rotate about any two angles, namely azimuthal and polar angle. However, in the 2D case only rotation with respect to a single axis is allowed. A basis functions with single rotation angle is simpler to be used than spherical harmonics.

Beside spherical harmonics, circular harmonics can be used to expand the intensitis as long as the random angles only have a single axis. The expression of the diffraction patterns in terms of circular harmonic expansion can be written

$$I(q, \phi) = \sum_m I_m(q) \exp(im\phi). \tag{2.18}$$

This is derived similarly as before, by substituting equation 2.18 into equation 2.9 and performing the average over all diffraction patterns. The new C_2 with respect to circular harmonics becomes

$$C_2(q, q'; \phi\phi') = \sum_m I_m^*(q) I_m(q') \exp(im(\phi - \phi')) \tag{2.19}$$

where $I_m(q)$ is the circular harmonic expansion coefficients of the diffracted intensity of a single particle. The right hand side of equation 2.19 is an exponential function. Multiplying both side with its inverse and integrating over all angles, will remove the dependence of the exponential function in the right hand side of the equation. Thus, a new quantity can be obtained, namely

$$B_m(q, q') \quad = \quad \int C_2(q, q', \Delta\phi) \exp(-im\Delta\phi) d\Delta\phi = I_m(q)^* I_m(q') \tag{2.20}$$

The information about the structure is contained in the quantity $I_m(q)$. The mag-

nitude of $I_m(q)$ is directly accessible by taking square root of the diagonal values of $B_m(q,q')$. For that reason, the phase of $I_m(q)$ is the only missing information to fully determine $I_m(q)$ from $B_m(q,q')$. After $I_m(q)$ is determined, the reconstruction of the intensity distribution of a single molecule can be found from equation 2.18.

Now after deriving $B_l(q,q')$ and $B_m(q,q')$, there is another quantity that is very important for reconstruction of structure of the molecule, namely two point angular triple correlations. Mathematically, it is defined by [32]

$$C_3(q,q',\phi,\phi') = \frac{1}{N_p}\sum_p I_p^2(q,\phi)I_p(q',\phi') \tag{2.21}$$

Using a simiilar derivation as before, the expansion coefficients in equation 2.8 are substituted into equation 2.21. The result of substitution is

$$\begin{aligned} C_3(q,q',\phi,\phi') = \frac{1}{N_p}\sum_p \sum_{l_1,l_2,l_3} \sum_{m_1,m_2,m_3} \sum_{m_1',m_2',m_3'} \\ \times I_{l_1m_1} D_{l_1m_1m_1'}(\omega) Y_{l_1m_1'}(\theta,\phi) \\ \times I_{l_2m_2} D_{l_2m_2m_2'}(\omega) Y_{l_2m_2'}(\theta,\phi) \\ \times I^*_{l_3m_3} D^*_{l_3m_3m_3'}(\omega') Y^*_{l_3m_3'}(\theta,\phi'). \end{aligned} \tag{2.22}$$

To simplify the above equation, these relations are substituted into the above equation:

$$D_{l_1 m_1 m_1'}(\omega) D_{l_2 m_2 m_2'}(\omega) = \sum_{L=|l_1-l_2|}^{l_1+l_2} \sum_{(M,M')=-L}^{L} (2L+1)(-1)^{M-M'} \times \begin{pmatrix} l_1 & l_2 & L \\ m_1 & m_2 & -M \end{pmatrix} \times \begin{pmatrix} l_1 & l_2 & L \\ m_1' & m_2' & -M' \end{pmatrix} D_{LMM'}(\omega), \tag{2.23}$$

$$Y_{l_1 m_1'}(\Omega) Y_{l_2 m_2'}(\Omega) = \sum_{\lambda=|l_1-l_2|}^{l_1+l_2} \sum_{\mu=\lambda}^{+\lambda} \left[\frac{(2l_1+1)(2l_2+1)(2l_1+1)}{4\pi} \right]^{1/2} \times \begin{pmatrix} l_1 & l_2 & \lambda \\ 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} l_1 & l_2 & \lambda \\ m_1' & m_2' & \mu \end{pmatrix} Y_{\lambda\mu}(\Omega), \tag{2.24}$$

$$\sum_{m_1' m_2'} \begin{pmatrix} l_1 & l_2 & \lambda \\ m_1' & m_2' & \mu \end{pmatrix} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1' & m_2' & -m_3' \end{pmatrix} = \frac{1}{2l_3+1} \delta_{\lambda l_3} \delta_{\mu m_3'}, \tag{2.25}$$

and

$$\sum_{m'3} Y_{l m_3'}(\theta, \phi) Y^*_{l m_3'}(\theta, \phi') = \frac{2l+1}{4\pi} P_l(\cos(\phi - \phi')) \tag{2.26}$$

where the quantities represented by the large parentheses are Wigner $3j$ symbols.

After substituting equations 2.23, 2.24, 2.25, and 2.26 into equation 2.22, the equation

2.22 becomes

$$C_3(q,q',\Delta\phi) = \sum_{l_1 l_2 l} \sum_{m_1 m_2 m} I_{l_1 m_1}(q) I_{l_2 m_2}(q) I^*_{lm}(q') P_l(\cos(\Delta\phi)) \\ \times (-1)^m (4\pi)^{-\frac{3}{2}} \begin{pmatrix} l_1 & l_2 & l \\ m_1 & m_2 & -m \end{pmatrix} \begin{pmatrix} l_1 & l_2 & l \\ 0 & 0 & 0 \end{pmatrix} \\ \times \left[(2l_1+1)(2l_2+1)(2l_1+1)\right]^{1/2}. \tag{2.27}$$

Additional relation is needed to invert the equation 2.27. The relation is the orthogonality of the Legendre polynomials, the mathematical expression is

$$\int_{-1}^{1} P_l(u) P_{l'}(u) du = \frac{2}{2l+1} \delta_{ll'}. \tag{2.28}$$

A new quantity $T_l(q,q')$ is obtained by applying the orthogonality of the Legendre polynomials into equation 2.27. The $T_l(q,q')$ can be written as

$$T_l(q,q') = \int C_3(q,q',\Delta\phi) P_l(\cos(\Delta\phi)) d(\Delta\phi) \frac{(2l+1)}{2} 4\pi \tag{2.29}$$

or theoretically can be calculated from

$$T_l(q,q') = \sum_{l_1,l_2,m_1,m_2,m} (-1)^m \left[\frac{(2l_1+1)(2l_2+1)(2l_1+1)}{4\pi}\right]^{1/2} \begin{pmatrix} l_1 & l_2 & l \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} l_1 & l_2 & l \\ m_1 & m_2 & -m \end{pmatrix} \\ I_{l_1,m_1}(q) I_{l_2,m_2}(q) I_{l,m}(q). \tag{2.30}$$

Apart from $B_l(q,q')$, the information about the structure of the molecule can be obtained from $T_l(q,q')$ as well. The C_3 is a quantity that is obtainable from experiment data as described in equation 2.21. Thus, $T_l(q,q')$ can be calculated from experimental

data as described in equation 2.29. As a result of that, $T_l(q,q')$ can be used to reveal the information about the structure of the molecule because it involves the summation over spherical harmonic expansion of the diffraction volume.

2.2.1 Independent Parameters

As stated before, $B_l(q,q')$ is one of the quantities measurable in the experiment. The objective of this method is to obtain the electron density from the $B_l(q,q')$. If the diffraction volume or intensity can be obtained from $B_l(q,q')$ then the diffraction volume can be phased using a phasing algorithm to get the electron density. Having said that, it is important to study ithe relationship between $B_l(q,q')$ and $I_{lm}(q)$.

For a given $B_l(q,q')$, $I_{lm}(q)$ cannot be determined uniquely. The reason of that is a new $I_{lm}(q)$ can be formed by multiplying it by orthogonal matrix.

$$I'_{lm}(q) = O^l_{mm'} I_{lm'}(q) \tag{2.31}$$

$$\text{where} \quad O^l_{mm'}(O^l_{mm'})^\dagger = 1. \tag{2.32}$$

In other words, if a matrix $O^l_{mm'}$ is unitary or orthogonal then the value of $B_l(q,q')$ is not affected by multiplication of any orthogonal matrix as shown below:

$$B'_l(q,q') = \sum_m I'_{lm}(q) I'^\dagger_{lm}(q') \tag{2.33}$$

$$B'_l(q,q') = \sum_m I_{lm}(q) O^l_{m'm''}(O^l_{m'm''})^\dagger I^\dagger_{lm}(q') \tag{2.34}$$

$$B'_l(q,q') = \sum_m I_{lm}(q) I^\dagger_{lm}(q')$$

$$B'_l(q,q') = B_l(q,q').$$

For each l, there are unitary matrices $O^l_{mm'}$ that contribute to the nonuniqueness of

$I_{lm}(q)$. The matrices $O^l_{mm'}$ have $2l+1$ rows and $2l+1$ columns. The total elements of the particular matrix is $(2l+1)^2$. However, not all elements are independent of each other because the matrix satisfis orthogonality.

From [34], an $n \times n$ orthogonal matrix has $\frac{n(n-1)}{2}$ independent elements. Since an $O^l_{mm'}$ has $(2l+1)$x$(2l+1)$ elements then the total independent elements for a particular l is $(2l+1)(l)$ elements.

Given the explanation above, the total elements is

$$\sum_{l=0,2,4,\ldots}^{l_{max}} (2l+1)(l). \tag{2.35}$$

2.3 Spherical Harmonics

2.3.1 Property of Spherical Harmonics

As mentioned in the previous section, the correlation method doesn't need to know the orientations of the individual diffraction patterns. It is very crucial to remove the angle dependence of the intensity since we want to recover the particle's structure. It is important that the selected function can be separated by its angle dependence and radius dependence. A set of functions that satisfies such a criterion are spherical harmonics.

Spherical harmonics are a series of special functions defined on the surface of sphere. It is defined in spherical coordinates represented by angles θ and ϕ. Spherical harmonics are characterized by two quantum numbers namely l and m. The quatum number m specifies how the function varies with respect to the azimuthal angle.

The definition of spherical harmonics is given by

$$Y_{lm}(\theta,\phi) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}} P_{lm}(\cos\theta) e^{im\phi} \tag{2.36}$$

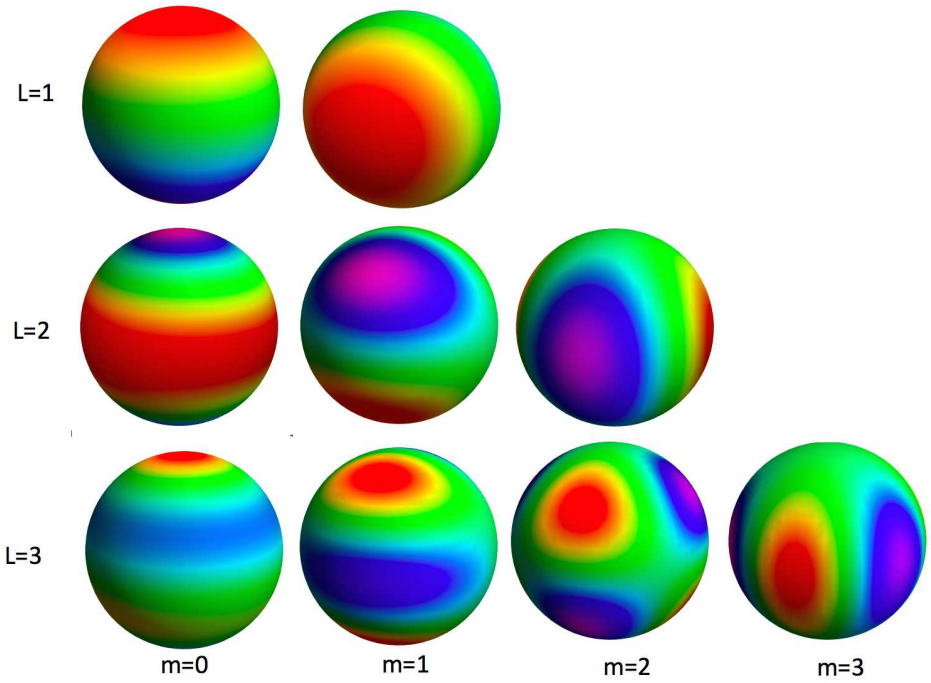


Figure 2.8: Example of plot of spherical harmonics with different quantum numbers

where the $P_{lm}(\cos\theta)$ are legendre polynomials. Legendre polynomial $P_{lm}(x)$ can be obtained using Rodrigues formula:

$$P_{lm}(x) = \frac{(-1)^m}{2^l l!}(1-x^2)^{m/2}\frac{d^{l+m}}{dx^{l+m}}\left[(x^2-1)^l\right]. \tag{2.37}$$

It is important to note that a spherical harmonic is a polynomial of trigonometric functionis. As in other polynomial expansions, a lower degree represents an approximation of the function and a higher degree contains information of how rapidly the function varies.

Spherical harmonics are a set of functions characterized by 2 quantum numbers. It is important to show the relation between those functions. Every single spherical harmonic function with different quantum numbers is orthogonal to each other. This relation may be summarized

$$\int Y_{lm}Y^*_{l'm'}d\Omega = \delta_{ll'}\delta_{mm'}. \tag{2.38}$$

where $\delta_{ll'}$ is ia Kronecker delta that is non zero if the two indices are the same.

The aim of this section is to characterize a symmetry in terms of spherical harmonic quantum numbers. In order to study rotational symmetry, a rotation operatoin the basis of the spherical harmonics is needed. One well known operator to rotate spherical harmonics is the Wigner D-matrix. The definition below shows how spherical harmonics are rotated,

$$Y_{lm}(\theta', \phi') = \sum_{m'} D^{l}_{mm'}(\alpha, \beta, \gamma) Y_{lm'}(\theta, \phi) \tag{2.39}$$

where θ, ϕ are with respect to original axes and the θ', ϕ' are with respect to axes rotated by Euler angles (α,β,γ). Elements of the Wigner D-matrix are calculated as follows:

$$D^{l}_{mm'}(\alpha, \beta, \gamma) = e^{im'\gamma} d^{j}_{mm'}(\beta) e^{-im\alpha} \tag{2.40}$$

and $d^{j}_{mm'}$ is calculated by applying summation:

$$d^{j}_{mm'}(\beta) = [(j+m')!(j-m)!(j+m)!(j-m)!]^{1/2} \tag{2.41}$$
$$\sum_{s} \frac{(-1)^{m'-m+s}}{(j+m-s)!s!(m'-m+s)!(j-m'-s)!} \cos(\frac{\beta}{2})^{2j+m-m'-2s} \sin(\frac{\beta}{2})^{m'-m+2s}]$$

2.3.2 Effect of Azimuthal Symmetry on Spherical Harmonics Expansion

It is very essential to discuss the azimuthal symmetry of the spherical harmonics. One important feature is how a coordinate transformation affects the expansion of spherical harmonics. It will be shown here how by rotating coordinates and by setting the z-axis as the center of symmetry, some components of spherical harmonics vanish.

In the figure 2.9, the z-axis is not aligned to the center of symmetry of the object. Even though the object is a cylinder, which has azimuthal symmetry, none of spherical harmonics components will be zero. The reason is that by rotating the object with respect

to z-axis the symmetry requirement is not satisfied. Having said that, the rotation of axes is very important to determine how the symmetry of an object affects the spherical harmonic expansion.

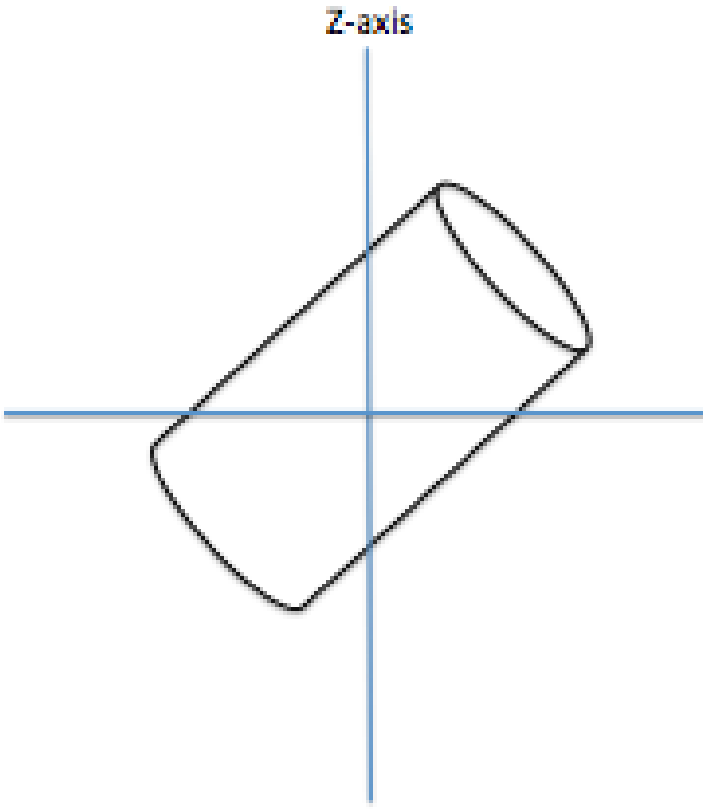


Figure 2.9: Rotation of z-axis doesn't reveal azimuthal symmetry

In figure 2.10, the z-axis is now aligned to the center of symmetry of object. There is no change in the appearance of the object by rotation with respect to z-axis. Since symmetry is found in this coordinate transformation, there is a pattern of allowed m quantum numbers in the spherical harmonic expansion. By equating spherical harmonics

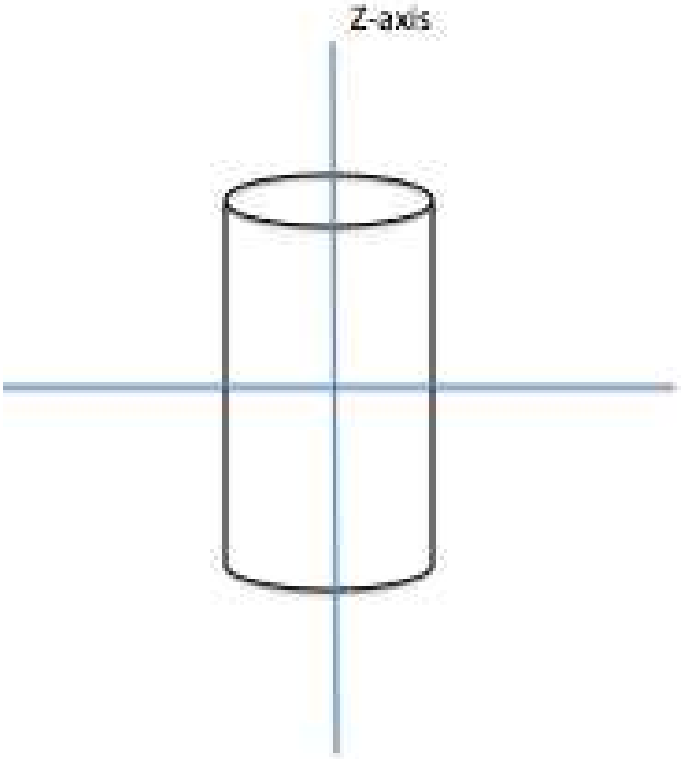


Figure 2.10: Rotation with respect to z-axis doesn't change the structure of object

before and after transformation,

$$\begin{aligned} Y_{lm}(\theta,\phi) &= Y_{lm}(\theta,\phi+\delta) \\ Y_{lm}(\theta,\phi) &= \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_{lm}(\cos\theta)e^{im(\phi+\delta)} \\ \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_{lm}(\cos\theta)e^{im(\phi)} &= \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_{lm}(\cos\theta)e^{im(\phi+\delta)} \\ e^{im(\phi)} &= e^{im(\phi+\delta)} \end{aligned} \tag{2.42}$$

$e^{im(\phi)} = e^{im(\phi+\delta)}$ is requirement to be satisfied if object has azimuthal symmetry with respect to z-axis. Since δ is any arbitrary angle, only $m = 0$ satisfis the equation as it is shown in table 2.2

m	$e^{im(\phi)} = e^{im(\phi+\delta)}$
m=0	1=1
m=1	$\cos(1\phi) + i\sin(1\phi) \neq \cos(1(\phi+\delta)) + i\sin(1(\phi+\delta))$
m=2	$\cos(2\phi) + i\sin(2\phi) \neq \cos(2(\phi+\delta)) + i\sin(2(\phi+\delta))$
m=3	$\cos(3\phi) + i\sin(3\phi) \neq \cos(3(\phi+\delta)) + i\sin(3(\phi+\delta))$
m=n	$\cos(n\phi) + i\sin(n\phi) \neq \cos(n(\phi+\delta)) + i\sin(n(\phi+\delta))$

Table 2.2: Only $m = 0$ satisfis azimuthal symmetry since δ is arbitrary angle

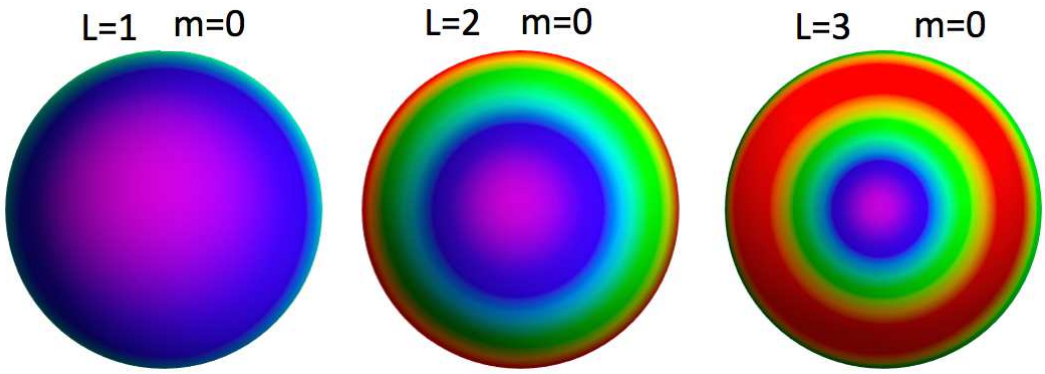


Figure 2.11: Plot of spherical harmonics with azimuthal symmetry

2.3.3 Effect of 4-fold symmetry on Spherical Harmonics Expansion

The behavior of spherical harmonics that have 4-fold symmetry will be thoroughly explained here. The reason 4-fold symmetry is important is that later the object under study is a K-channel protein that satisfies 4-fold symmetry. Studying which expansion vanishes for given particular m quantum number enables one to determine if the object under study has 4-fold symmetry.

As in the case of azimuthal symmetry, 4-fold symmetry is the rotational symmetryelement with respect to the z axis. The spherical harmonic axis can be arbitrary rotated, by setting the center of symmetry as z-axis, the selection rule will appear as a result of the symmetry of the object.

Figure 2.12 is example of an object which has 4-fold symmetry and the center of

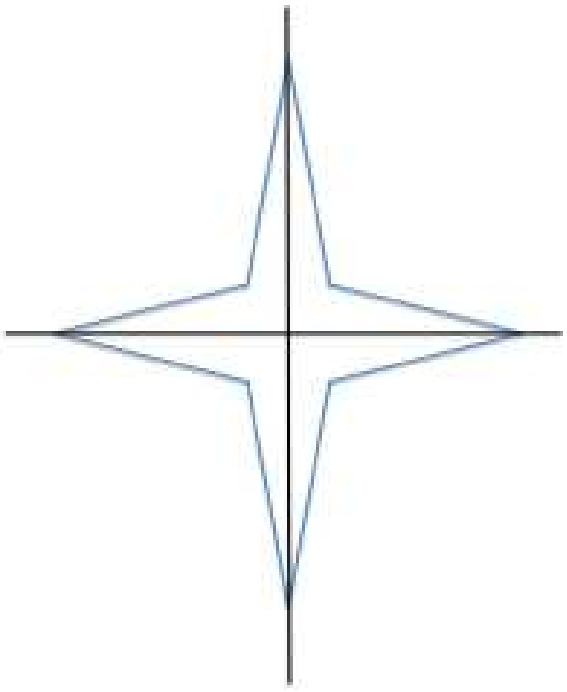


Figure 2.12: Top view of object with 4-fold symmetry, rotation by 90^0 doesn't change the appearance of the object

symmetry is aligned with the z-axis. Rotation of angle 90^0 or $\pi/2$ doesn't change the structure of the object. By equating spherical harmonics with the rotated one, one can find the quantum number that satisfies 4-fold symmetry.

$$\begin{aligned} Y_{lm}(\theta,\phi) &= Y_{lm}(\theta,\phi+\frac{\pi}{2}) \\ Y_{lm}(\theta,\phi) &= \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_{lm}(\cos\theta)e^{im(\phi+\pi/2)} \\ \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_{lm}(\cos\theta)e^{im(\phi)} &= \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_{lm}(\cos\theta)e^{im(\phi+\pi/2)} \\ e^{im(\phi)} &= e^{im(\phi+\pi/2)} \end{aligned} \tag{2.43}$$

$e^{im(\phi)} = e^{im(\phi+/pi/2)}$ is the requirement to be satisfied if the object has 4-fold symmetry with respect to z-axis. Table 2.3 shows what quantum number persist if the object has 4-fold symmetry.

m	$e^{im(\phi)} = e^{im(\phi+\pi/2)}$
m=0	1=1
m=1	$\cos(1\phi) + i\sin(1\phi) \neq \cos(1(\phi+\pi/2)) + i\sin(1(\phi+\pi/2))$
m=2	$\cos(2\phi) + i\sin(2\phi) \neq \cos(2(\phi+\pi/2)) + i\sin(2(\phi+\pi/2))$
m=3	$\cos(3\phi) + i\sin(3\phi) \neq \cos(3(\phi+\pi/2)) + i\sin(3(\phi+\pi/2))$
m=4	$\cos(4\phi) + i\sin(4\phi) = \cos(4(\phi+\pi/2)) + i\sin(4(\phi+\pi/2))$
m=5	$\cos(5\phi) + i\sin(5\phi) \neq \cos(5(\phi+\pi/2)) + i\sin(5(\phi+\pi/2))$
m=6	$\cos(6\phi) + i\sin(6\phi) \neq \cos(6(\phi+\pi/2)) + i\sin(6(\phi+\pi/2))$
m=7	$\cos(7\phi) + i\sin(7\phi) \neq \cos(7(\phi+\pi/2)) + i\sin(7(\phi+\pi/2))$
m=8	$\cos(8\phi) + i\sin(8\phi) = \cos(8(\phi+\pi/2)) + i\sin(8(\phi+\pi/2))$

Table 2.3: Only $m = 4n$, where n is integer, satisfy 4-fold symmetry

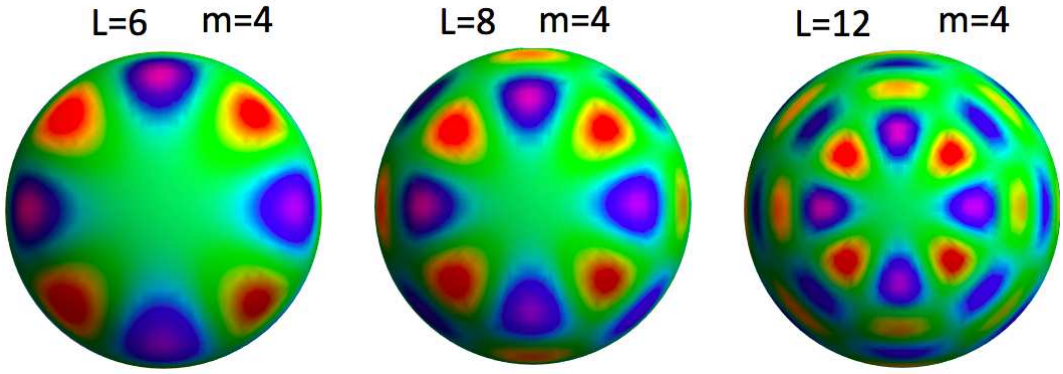


Figure 2.13: Plot of spherical harmonics with 4-fold symmetry

2.3.4 Effect of Icosahedral symmetry on Spherical Harmonics Expansion

The behavior of spherical harmonics that have icosahedral symmetry will be thoroughly explained here. Previously, the symmetry under study is based on rotation of one axis only and the pattern involves only the m quantum number. More complicated pattern will arise and quantum number in both m and l are necessary. One of symmetries which has more than one rotational axis is icosahedral symmetry. Studying which expansion vanishes for given particular m quantum number enable one to determine if the object under study has 4-fold symmetry.

Different than azimuthal and 4-fold symmetry, icosahedral symmetry has 3 rotational

axes. They are 5-fold, 3-fold and 2-fold axes. The z-axis can be chosen arbitrary, by setting the center of symmetry as the 5-fold axis unique selection rule will appear as a result of the symmetry of the object. Based on icosahedral selection rule[24], I_{lm} is nonzero when l satisfy.

$$l = 6p + 10q \tag{2.44}$$

where p and q in integer

and m quantum numbers are

$$m = ..., -10, -5, 0, 5, 10, ... \tag{2.45}$$

when of 5-fold axis is taken as the z-axis. A function can be constructed from a linear

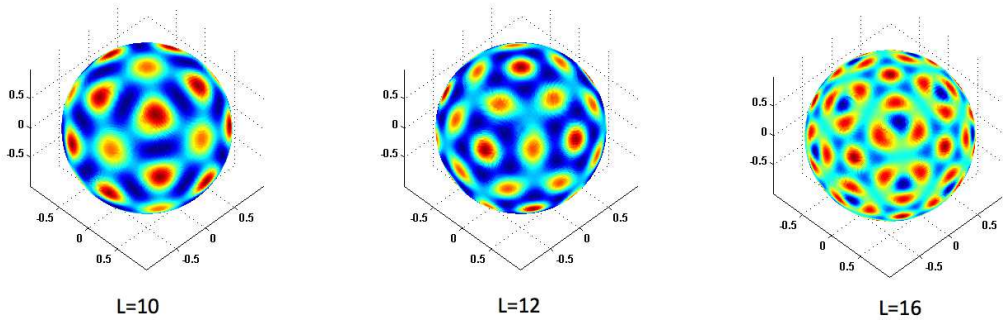


Figure 2.14: Plot of spherical harmonics with icosahedral symmetry

combination of spherical harmonics. In order the function satisfies icosahedral symmetry only spherical harmonics which satisfies the selection rule are taken into the linear combination. In equation 2.46, $J_l(\theta, \phi)$ is an icosahedral harmonic which consist of a linear combination of spherical harmonics. By summing over all m, icosahedral harmonics only depend on the l quantum number.

The factor a_{lm} cannot be arbitrary because equation 2.46 must satisfy icosahaderal symmetry[23]. Table 2.4 shows values of a_{lm} for different combination of l and m. Icosa-

hedral harmonics are defined as linear combination of spherical harmonics which satisfy icosahedral symmetry:

l m	0	5	10	15	20
0	1.0				
6	0.531085	0.847318			
10	0.265539	-0.846143	0.462094		
12	0.454749	0.469992	0.75613		
16	0.334300	-0.493693	-0.634406	0.491975	
18	0.399497	0.450611	0.360958	0.712083	
20	0.077539	-0.460748	0.747888	-0.231074	0.411056

Table 2.4: Coefficient's a_{lm} of spherical harmonics to convert into icosahedral harmonics [17]

$$J_l(\theta, \phi) = \sum_m a_{lm} Y_{lm}(\theta, \phi) \tag{2.46}$$

From table 2.4, I_{lm} is nonzero if m is a multiple 5. By looking at equation 2.46, for a particular l, I_{lm} is not independent if the object has icosahedral symmetry. The spherical harmonics expansion of an icosahedral object only depends on values of l, given the l the values of m are determined by symmetry and are tabulated. In other words, for icosahedral object there is one independent parameter of icosahedral harmonics for each I_{lm}.

2.4 Symmetry of Angular Correlations

2.4.1 Rotation of Data Points

This section explains the relation between a rotation matrices and data points. It will be shown that redundancy or lowest number of independent parameter can be found by applying a particular rotation matrix on data points.

An orthogonal transformation is a linear transformation that preserves the dot products of vectors. The length or radius of the vectors are not changed by applying an orthogonal transformation. Even more, the angle between two vectors is preserved. Applying an orthogonal transformation on the coordinate axes will result in rotation, reflection, or inversion of axes. Mathematically, an orthogonal transformation is represented as a rotation matrix. The basic theory of orthogonal transformations and rotation matrices is described in this section.

(a) Original axis

(b) Rotated axis

Figure 2.15: Any point can be described in transformed axis

Figure 2.15 illustrates that any vector can be described in terms of any of the axes. As long as the relation of the new to the old axes is caused by an orthogonal transformation, the effect on the vectors joining any of the data pints to the origin is only a rotation, preserving the radius or length of the vectors. Throughought all rotations, there will always be an axis or direction in which one particular axis will have a smallest component as displayed in the figure 2.16, By knowing that axis, it can be used to reduce the dimension of the data without losing essential information because the axis with the smallest component has the least information.

A rotation matrix can be used to indicate whether there is redundant information in a data set. A redundancy means there is a different way of representing the data with a lower number of independent parameters. Illustrated in figure 2.17, the data points

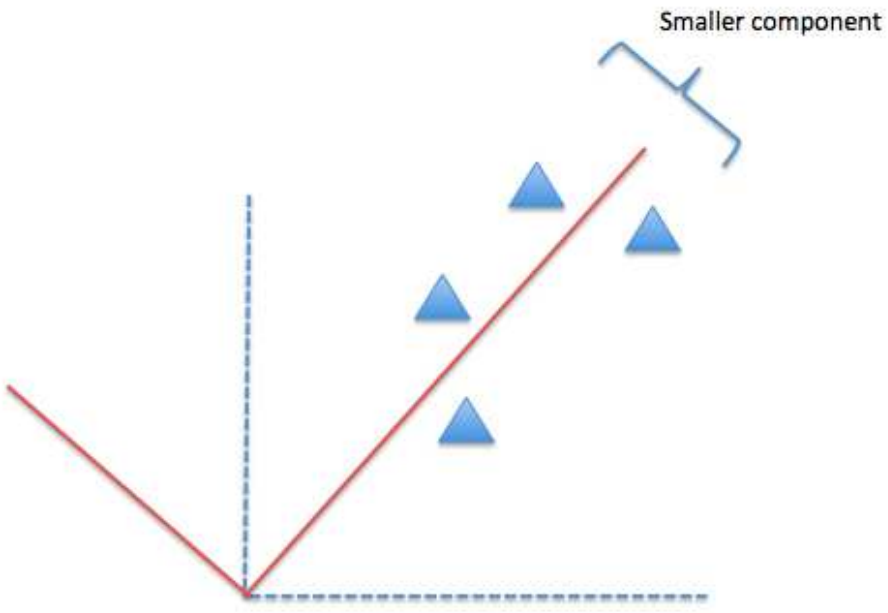


Figure 2.16: Red is the axis which has maximum variance in one direction and minimum component in another one

are represented by two different ways, using blue axes the data are specified using two parameters whereas using the red axis the data are specified with one parameter. If the data contains redundant information, then the number of independent parameters can be reduced by rotating the axes. Figure 2.17 shows that redundancy in 2D occurs due to the data lying in the same line. Generalizing into higher dimension, the redundancy occurs when the data lies in either a line, a plane, or a hyperplane.

An essential property to find the lowest number of independent parameters is by knowing that the dot products between the data sets are enough to reveal the redundancies. From figure 2.17, all data points have the same angle from each other. Since the angle is obtainable from the dot product, by constructing a matrix of dot products, a pattern appears indicating whether there is a redundancy inside data sets. The redundancy in the 2D case is very simple; if the angles are the same for the all data set, then there is redundancy. To reveal the redundancy in higher order, a different sophisticated method

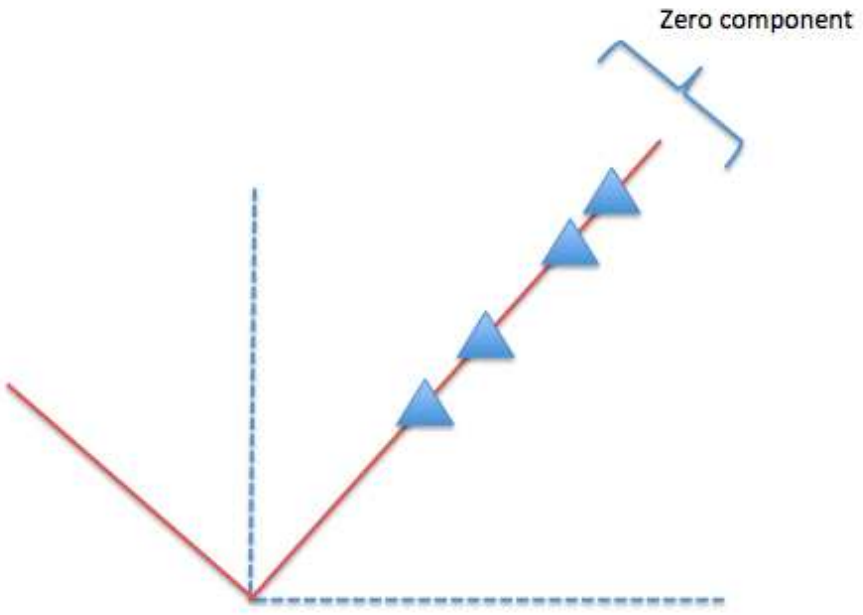


Figure 2.17: In red axis, data can be specified with one parameter only

is needed, because a common pattern in the dot product is not easily observable.

One of the methods to reveal redundancy of the data in higher dimension is principal component analysis. The next section will explain principal component analysis and how it can be used to determine symmetry of particles only from correlation amongst the data.

2.4.2 Principal Component Analysis

Principal component analysis (PCA) is an established method to reduce the dimension of a set of vectors. PCA uses orthogonal transformation is to convert a set of vectors into a set of new vectors. The determination of the orthogonal transformation are defined in such a way that one axis will have the largest variance and another axis will have the smallest component. In addition to that, PCA can be used to find the lowest number of independent parameters in a data set, which is the purpose of this section.

Another important point is that the rotation of an axis can be used to find the lowest

number of independent parameters in the data set. A set of vectors in any arbitrary independent axis can be described as

$$\begin{aligned}
\vec{V_1} &= \begin{bmatrix} \upsilon_{11} & \upsilon_{12} & \dots & \upsilon_{1m} \end{bmatrix} \\
\vec{V_2} &= \begin{bmatrix} \upsilon_{21} & \upsilon_{22} & \dots & \upsilon_{2m} \end{bmatrix} \\
\vec{V_3} &= \begin{bmatrix} \upsilon_{31} & \upsilon_{32} & \dots & \upsilon_{3m} \end{bmatrix} \\
\vec{V_n} &= \begin{bmatrix} \upsilon_{n1} & \upsilon_{n2} & \dots & \upsilon_{nm} \end{bmatrix}
\end{aligned} \tag{2.47}$$

.

To represent this set of vectors as a set of new data, a matrix can be formed by arranging each row as new independent vector and each column as an independent component. The matrix is

$$\mathbf{V} = \begin{bmatrix} \upsilon_{11} & \upsilon_{12} & \dots & \upsilon_{1m} \\ \upsilon_{21} & \upsilon_{22} & \dots & \upsilon_{2m} \\ \upsilon_{31} & \upsilon_{32} & \dots & \upsilon_{3m} \\ \vdots & \vdots & & \\ \upsilon_{n1} & \upsilon_{n2} & \dots & \upsilon_{nm} \end{bmatrix}. \tag{2.48}$$

From that new definition, a new quantity called a covariance matrix defined as

$$\mathbf{C} = \mathbf{V^t V}. \tag{2.49}$$

Based on PCA, the first eigenvector of the covariance matrix is the direction of the maximum variance or the minimum residual component. In addition to that, finding the lowest independent parameters to describe the system can be found from counting the nonzero eigenvalues of the covariance matrix. In addition to that, calculating eigenvectors

and eigenvalues can be done by using singular value decomposition (SVD). Mathematically, the SVD is

$$[u\, s\, v] = \mathrm{SVD}(\mathbf{V^t V}) \tag{2.50}$$

where u is a matrix composed of independent eigenvectors and s is a matrix consisting of eigenvalues of the covariance matrix.

Neither $\mathbf{V^t V}$ nor $\mathbf{V}$ are available from the angular correlation data. Only the matrix of dot product is available. It will be shown below that matrix of dot product will have eigenvalues equal to the eigenvalues of the covariance matrix.

The matrix of the dot product is defined as

$$\mathbf{M_d} = \begin{bmatrix} \vec{V_1}\cdot\vec{V_1} & \vec{V_1}\cdot\vec{V_2} & \dots & \vec{V_1}\cdot\vec{V_m} \\ \vec{V_2}\cdot\vec{V_1} & \vec{V_2}\cdot\vec{V_2} & \dots & \vec{V_2}\cdot\vec{V_m} \\ \vec{V_3}\cdot\vec{V_1} & \vec{V_3}\cdot\vec{V_2} & \dots & \vec{V_3}\cdot\vec{V_m} \\ \vdots & \vdots & & \\ \vec{V_n}\cdot\vec{V_1} & \vec{V_n}\cdot\vec{V_2} & \dots & \vec{V_n}\cdot\vec{V_m} \end{bmatrix} \tag{2.51}$$

$$= \vec{V}\vec{V}^t. \tag{2.52}$$

Having defined the matrix of the dot product, its eigenvalue can be found from the covarianve matrix or covariance matrix's eigenvalue can be found from the matrix of the

dot product. The proof is shown below:

$$\begin{aligned} \mathbf{M_d}\,\nu &= \lambda\,\nu \\ \vec{V}\,\vec{V}^t\,\nu &= \lambda\,\nu \\ \vec{V}^t\,\vec{V}\,\vec{V}^t\,\nu &= \lambda\,\vec{V}^t\,\nu \\ \vec{V}^t\,\vec{V}\,\mu &= \lambda\,\mu \end{aligned} \tag{2.53}$$

In conclusion, $\vec{V}\,\vec{V}^t$ and $\vec{V}^t\,\vec{V}$ have equal eigenvalues but different eigenvectors. The eigenvalue that give the lowest independent component is the eigenvalue of the covariance matrix. Hence, the eigenvalue of the covariance matrix can be easily be calculated by finding the eigenvalueis of the matrix of the dot product.

2.4.3 Matrix Correlation

As described in section 2.2, the final expression obtained from the correlation data is

$$B_l(q, q') = \sum_m I_{lm}(q) I_{lm}(q')^*. \tag{2.54}$$

It is important to note that $B_l(q, q')$ is a form of dot product depending of how one constructs vectors from $I_{lm}(q)$. Since every $I_{lm}(q)$ comes from a spherical harmonic decomposition, every element is independent. A set of new vectors can be constructed where the values of the m's correspond to the components and q's, and values of the angular momentum quantum numbers l's specify the vectors. As an example, the vectors constructed

from this definition are

$$
\begin{aligned}
I_l(q_1) &= \begin{bmatrix} I_{l(-l)}(q_1) & I_{l(-l+1)}(q_1) & \dots & I_{l(l)}(q_1) \end{bmatrix} \\
I_l(q_2) &= \begin{bmatrix} I_{l(-l)}(q_2) & I_{l(-l+1)}(q_2) & \dots & I_{l(l)}(q_2) \end{bmatrix} \\
I_l(q_3) &= \begin{bmatrix} I_{l(-l)}(q_3) & I_{l(-l+1)}(q_3) & \dots & I_{l(l)}(q_3) \end{bmatrix} \\
I_l(q_n) &= \begin{bmatrix} I_{l(-l)}(q_n) & I_{l(-l+1)}(q_n) & \dots & I_{l(l)}(q_n) \end{bmatrix}.
\end{aligned}
\tag{2.55}
$$

The expression $\sum_m I_{lm}(q) I_{lm}(q')^*$ is equivalent to $\langle I_l(q), I_l(q') \rangle$ that is the dot product of $I_l(q)$. Now with that definition, $B_l(q, q')$ is a dot product of vectors $I_{lm}(q)$.

After confirming that $B_l(q, q')$ is the dot product of the vector $I_{lm}(q)$, a new matrix must be constructed in order to be used in PCA. There are an infinite number of possible ways to construct a matrix from a set of vectors. In order to be used in PCA, the matrix of dot products is constructed by arranging all possible dot products of different vectors

or q points. Below is how matrix is constructed

$$\mathbf{B^l_{qq'}} = \begin{bmatrix} \langle I_l(q_1), I_l(q_1)\rangle & \langle I_l(q_1), I_l(q_2)\rangle & \dots & \langle I_l(q_1), I_l(q_m)\rangle \\ \\ \langle I_l(q_2), I_l(q_1)\rangle & \langle I_l(q_2), I_l(q_2)\rangle & \dots & \langle I_l(q_2), I_l(q_m)\rangle \\ \\ \langle I_l(q_3), I_l(q_1)\rangle & \langle I_l(q_3), I_l(q_2)\rangle & \dots & \langle I_l(q_3), I_l(q_m)\rangle \\ \vdots & \vdots & \vdots & \\ \langle I_l(q_n), I_l(q_1)\rangle & \langle I_l(q_n), I_l(q_2)\rangle & \dots & \langle I_l(q_n), I_l(q_m)\rangle \end{bmatrix} \tag{2.56}$$

$$= \begin{bmatrix} B_l(q_1, q_1) & B_l(q_1, q_2) & \dots & B_l(q_1, q_m) \\ \\ B_l(q_2, q_1) & B_l(q_2, q_2) & \dots & B_l(q_2, q_m) \\ \\ B_l(q_3, q_1) & B_l(q_3, q_2) & \dots & B_l(q_3, q_m) \\ \vdots & \vdots & \vdots & \\ B_l(q_n, q_1) & B_l(q_n, q_2) & \dots & B_l(q_n, q_m) \end{bmatrix} \tag{2.57}$$

All elements of matrix $\mathbf{B_{qq'}}$ are obtainable from experiment according to eq 2.15. The matrix satisfies the requirement to be a matrix of dot products as given in eq. 2.54. The singular values of the matrix contain the information about the redundancy of the data. By counting the number of nonzero singular value, those number can be used to describe the redundancy in vector $I_{lm}(q)$.

The number of significant nonzero singular values represent total number parameters to describe the data. Only nonzero singular values contribute to the independent parameters. The nonsignificant or zero singular values denote the number of redundant parameters. By comparing how many significant or nonzero, nonsignificant or zero, and total singular values, those information later is essential to predict the symmetry of the

particle.

Chapter 3

Result

3.1 Dependence of the Number of m values on Symmetry

3.1.1 Azimuthal Pattern

Given in section 2.4.1, there will be a particular pattern of nonzero singular values depending on the symmetry of the object. In the section 2.3.2, it is explained that the azimuthal spherical harmonics components can be described by only one m value for each l or in other words only $m = 0$ is nonzero. If the object has azimuthal symmetry then the matrix $B_l(q, q')$ will only have one non zero significant singular value.

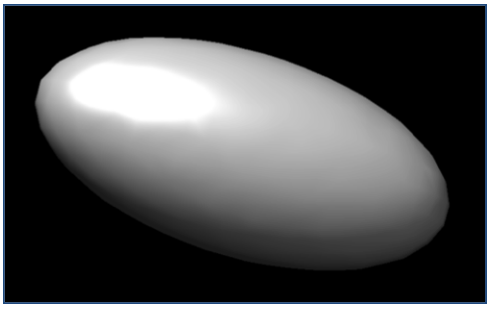


Figure 3.1: Model which has azimuthal symmetry

Figure 3.1 is a model that is used to calculate $B_l(q, q')$. The model is an ellipsoid which satisfies pure azimuthal symmetry, therefore $B_l(q, q')$ inherently contains azimuthal symmetry. By taking the SVD of the $B_l(q, q')$, the redundancy will be revealed and can be used to deduce the symmetry of object.

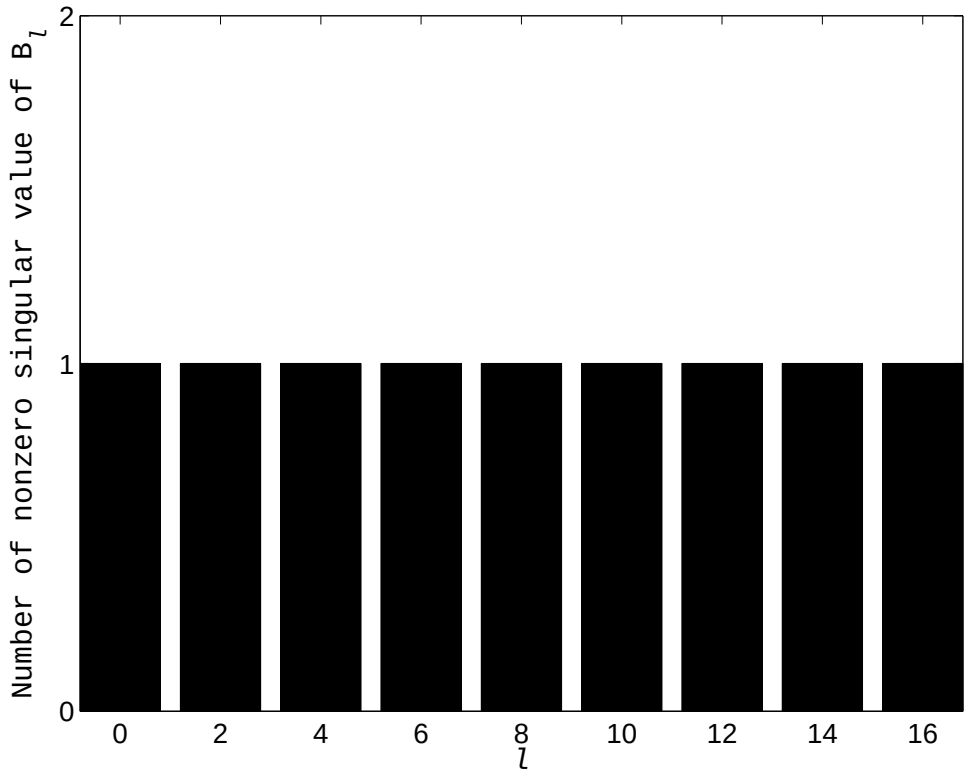


Figure 3.2: Total number of nonzero singular values vs angular momentum

l\m	-8	-6	-4	-2	0	2	4	6	8
0	0	0	0	0	$I_{lm} \neq 0$	0	0	0	0
2	0	0	0	0	$I_{lm} \neq 0$	0	0	0	0
4	0	0	0	0	$I_{lm} \neq 0$	0	0	0	0
6	0	0	0	0	$I_{lm} \neq 0$	0	0	0	0
8	0	0	0	0	$I_{lm} \neq 0$	0	0	0	0

Figure 3.3: Table of nonzero I_{lm} for azimuthal symmetry

The graph on figure 3.2 shows the behavior of the singular value of the object which has azimuthal symmetry. It shows that only one nonzero singular value for each l, therefore

it matches the behavior of I_{lm} in which only $m = 0$ does not vanish. Having said that, the SVD of $B_l(q, q')$ can be used to predict if the object satisfies azimuthal symmetry.

3.1.2 4-fold Pattern

The example given here is for the object that has 4-fold symmetry. However in general it can be extended to any n-fold symmetry without losing of uniqueness. In the section 2.3.3, it is explained that if a 4-fold symmetry exists then the component of spherical harmonics, which the m's are a multiple value of 4, will be nonzero. As consequence of that, the number of nonzero singular values of the matrix $B_l(q, q')$ has a pattern that matches with the total number of nonzero components in spherical harmonics expansion.

Figure 3.4: K-channel protein has 4-fold symmetry

The figure 3.4 shows 4-fold symmetric model that is used to calculate $B_l(q, q')$. Because the model has 4-fold symmetry, the $B_l(q, q')$ inherently contain information about the 4-fold symmetry. By taking the SVD of the $B_l(q, q')$, the redundancy will be revealed

and can be used to deduce the symmetry of object. From figure 3.5 and table 3.6, there is a matching pattern. By comparing them, for $l = 2$ there is one nonzero singular value and there is only one nonzero I_{lm} that is when $m = 0$. Another example is for $l = 6$, there are 3 singular values from figure 3.5 and from table 3.6 there are 3 nonzero I_{lm} that is when $m = -4, 0, 4$. If unknown structure give the same behavior as in graph 3.5 then one can conclude it has 4-fold symmetry.

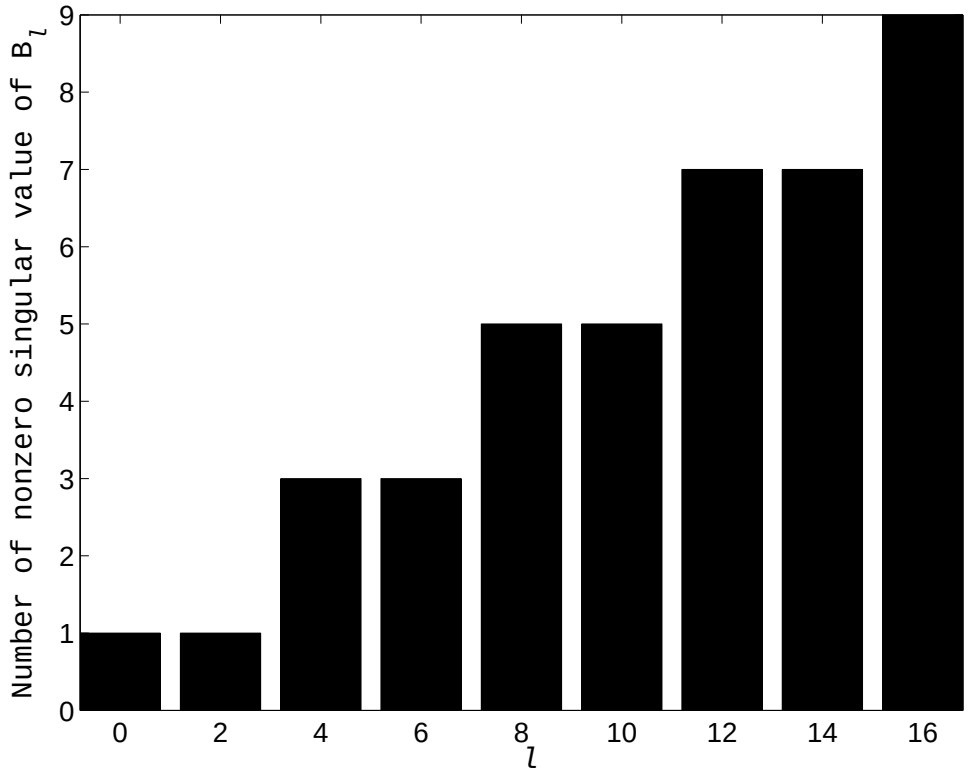


Figure 3.5: Total number of nonzero singular values vs angular momentum

l\m	-8	-6	-4	-2	0	2	4	6	8
0	0	0	0	0	$I_{lm} \neq 0$	0	0	0	0
2	0	0	0	0	$I_{lm} \neq 0$	0	0	0	0
4	0	0	$I_{lm} \neq 0$	0	$I_{lm} \neq 0$	0	$I_{lm} \neq 0$	0	0
6	0	0	$I_{lm} \neq 0$	0	$I_{lm} \neq 0$	0	$I_{lm} \neq 0$	0	0
8	$I_{lm} \neq 0$	0	$I_{lm} \neq 0$	0	$I_{lm} \neq 0$	0	$I_{lm} \neq 0$	0	$I_{lm} \neq 0$

Figure 3.6: Table of nonzero I_{lm} for 4-fold symmetry

3.1.3 Icosahedral Pattern

One of the important symmetries to be demonstrated in this section is icosahedral symmetry. The previous section explains that there is only one singular value of $B_l(q, q')$ if the object has icosahedral symmetry. In addition to that, there is a selection rule of l's for icosahedral harmonics. In order to predict the icosahedral symmetry both the selection rule and the singular value has to be satisfied.

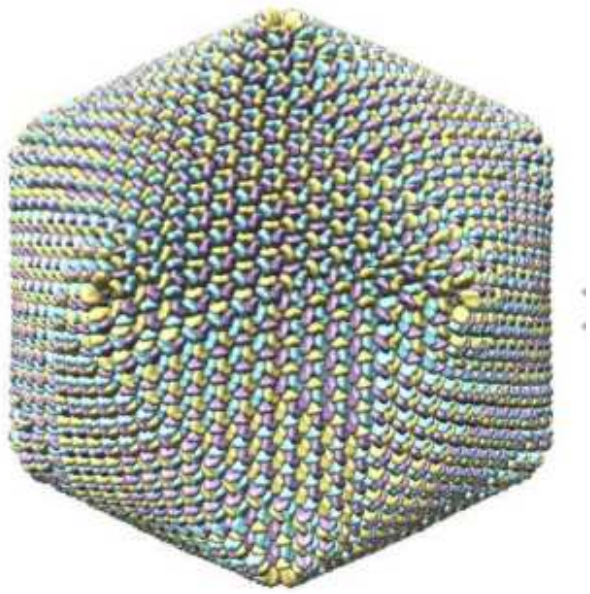


Figure 3.7: PBCV from pdb(1m4x) is used as model that has icosahedral symmetry [25]

The figure 3.7 shows a model that is used to calculate $B_l(q, q')$. The model is a virus that satisfies pure icosahedral symmetry. Thus, $B_l(q, q')$ inherently contains the icosahedral symmetry. By taking SVD of $B_l(q, q')$, the redundancy will be revealed and can be used to deduce the symmetry of the object.

The figure 3.8 is a graph calculated from the SVD of $B_l(q, q')$. It is obviously seen that it follows the icosahedral selection rule and at the same time has one singular value for each l. If unknown structure gives behavior as in graph 3.8 then one can conclude it has icosahedral symmetry.

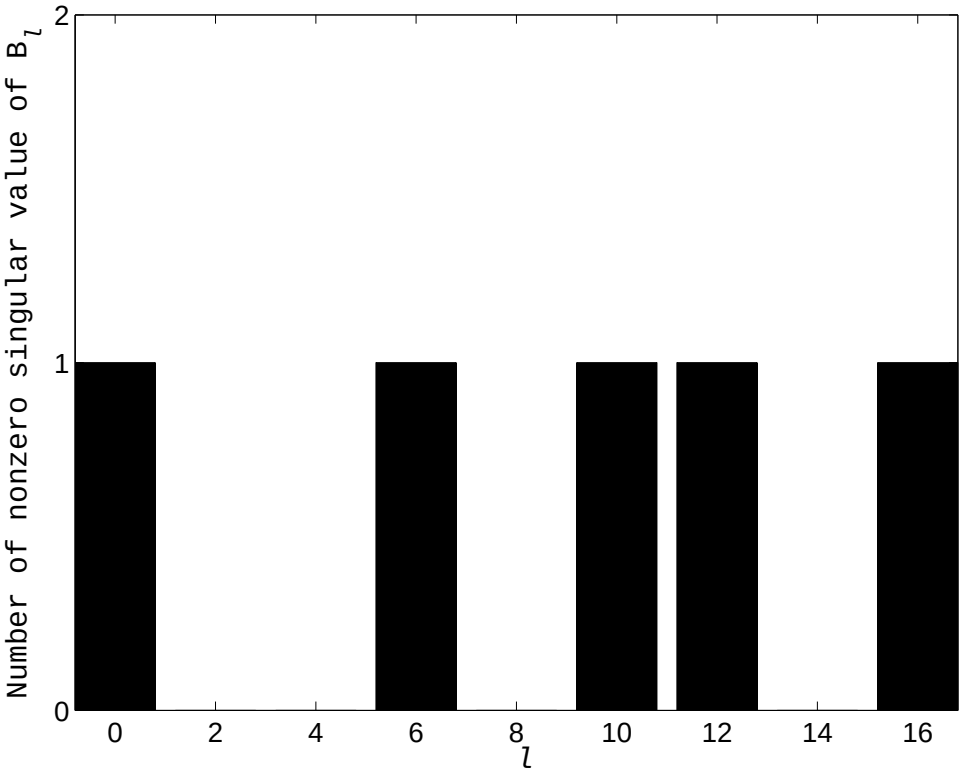


Figure 3.8: Total number of nonzero singular values vs angular momentum

3.1.4 Asymmetric Pattern

In this section distinguishing asymmetric property is demonstrated. The previous section explains that there will be $2l + 1$ singular values if the object under study does not have any particular symmetry. The number of singular values is equal to total number of m componentis of the spherical harmonics expansion. The figure 3.9 is a model that is used to calculate $B_l(q, q')$. The model is pyp protein, which does not have symmetry at all, therefore $B_l(q, q')$ inherently contain asymmetric property. From figure 3.10, for every l's there are $2L + 1$ singular values. As already mentioned before, $2l + 1$ singular values represent asymmetric pattern. If an unknown structure give a behavior as in graph 3.10 then one can conclude that it has asymmetry property.

This asymmetric pattern can be used to inspect whether $B_l(q, q')$ is a form of a dot product. Because no shape can be more asymmetrical than asymmetric shape, no shape

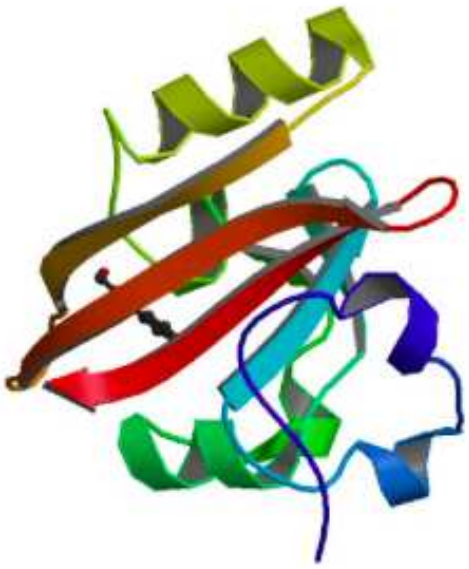


Figure 3.9: Photoactive yellow protein from pdb(2phy) is used as model

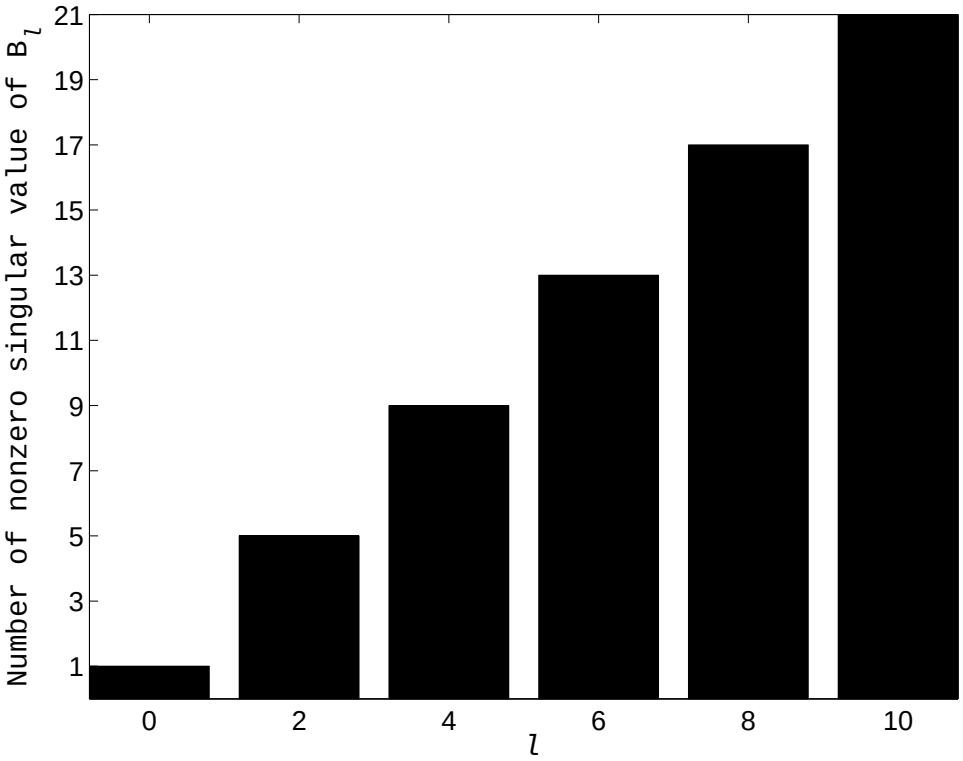


Figure 3.10: Total number of nonzero singular values vs angular momentum

can have more singular values than an asymmetric shape. In other words, $2l+1$ is the highest number of singular values and all shapes should have number of singular values less than or equal $2l+1$. If an SVD on the $B_l(q,q')$ has more singular values than $2l+1$ then the $B_l(q,q')$ is not a form of dot product or convergence is not reached.

3.1.5 Inversion Symmetry

This section describes one of the limits of the methods (PCA and selection rule) by giving an analysis of applying inversion symmetry. Because the available data from the experiment is a collection of diffraction patterns, all analyses initially have to be done in reciprocal space. It applies also to the determination of symmetry because the actual symmetry being determined is the symmetry of the diffraction volume. The symmetry of the structure is deduced from the symmetry of the diffraction volume because their relation is a Fourier transform and an operation of Fourier transform preserves the symmetry. As a result of that, PCA and the selection rule are used to determine the symmetry of the molecule implicitly through the determination of the symmetry in the diffraction volume.

If a molecule has inversion symmetry then each atom can be moved along inversion center to a point of equal distance without changing the whole shape of the molecule. The operation of inversion symmetry is changing each point (x,y,z) to $(-x,-y,-z)$ where $(0,0)$ is the inversion center. Any function that has inversion symmetry is invariant under the operation of inversion symmetry. Mathematically, it is described as

$$f(x,y,z) = f(-x,-y,-z). \tag{3.1}$$

Spherical harmonics have a distinct property to reveal the existence of inversion symmetry in a function. The analysis comes from the property of spherical harmonics where it is multiplication between an exponential and a Legendre polynomial. The inversion symmetry in polar coordinate is by tranforming $\theta \to \pi - \theta$ and $\phi \to \phi + \pi$. Mathemati-

cally, under the invariance of inversion symmetry from equation 3.1, spherical harmonics follow this property where

$$\begin{aligned} Y_{lm}(\theta,\phi) &= Y(\pi-\theta,\phi+\pi) \\ &\propto P_{lm}(\cos(\pi-\theta))\exp(im(\phi+\pi)) \\ &\propto (-1)^{m+l}P_{lm}(\cos(\theta))(-1)^m\exp(im(\phi)) \\ &\propto (-1)^l Y_{lm}(\theta,\phi) \end{aligned} \tag{3.2}$$

The expression in equation 3.2 always holds true for all components. Similarly, the spherical harmonics expansion ($I_{lm}(q)$) also follows the relation in equation 3.2. From equation 3.2, it is obvious that if the original function has inversion symmetry then its spherical harmonics expansion are zero for odd l and non-zero for even l.

In the previous discussion of symmetry, only even l are shown because all odd l are equal to zero. In other words, even though the original model doesnt have inversion symmetry. their $I_{lm}(q)$ have inversion symmetry (the $I_{lm}(q)$ are defined as coefficients of a spherical harmonic expansion of a diffraction volume). The diffraction volume is an absolute square of a structure factor, in which the structure factor is the result of a Fourier transfom of an electron density. Because the electron density is a quantity, which is described by real number, the following relations hold true:

$$\begin{aligned} A(\vec{q}) &= \int \rho(\vec{r})\exp(2\pi i\vec{q}\cdot\vec{r}) \\ A^*(\vec{q}) &= \int \rho(\vec{r})^*\exp(-2\pi i\vec{q}\cdot\vec{r}) \\ A^*(\vec{q}) &= \int \rho(\vec{r})\exp(-2\pi i\vec{q}\cdot\vec{r}) \\ A^*(-\vec{q}) &= A(\vec{q}) \\ |A^*(-\vec{q})|^2 &= |A(\vec{q})|^2 \\ I(-\vec{q}) &= I(\vec{q}). \end{aligned} \tag{3.3}$$

The relation is called Friedel's law. In other words, all diffraction volumes will always have inversion symmetry even though the electron density doesnt have inversion symmetry.

In conclusion, the inversion symmetry cannot be distinguished by using PCA nor by using selection rule. Another key point is that all electron density are described by real numbers (not complex numbers) and always have inversion symmetry in its diffraction volume regardless of whether the original electron density has inversion symmetry or not. By knowing only the diffraction volume, it is not sufficient to deduce if there is such symmetry in the electron density, therefore the inversion symmetry cannot be determined by the method explained above.

3.1.6 Experimental Data

This section mainly explains how to calculate $B_l(q, q')$ from experiment data and the convergence of $B_l(q, q')$. For this reason, the experimential diffraction patterns of nanorice were used to calculate $B_l(q, q')$. The diffraction patterns are available online and can be downloaded from cxidb.org.

The initial step that I did to analyze the experiment data was to separate good data from bad data. Figure 3.11 are some examples of the good data. Because it is known that the molecule is nanorice, it is expected to be close to an ellipsoid. In addition to that, the diffraction patterns of ellipsoid can be easily identified. Thus, those several patterns in the figure 3.11 shows the behavior where the diffracted molecules have the ellipsoidal property. Currently, by checking the data visually, I collected 200 good diffraction patterns.

The exclusion of bad data is simpler than finding a good ones. If the diffraction patterns does not contain signal then they are bad data. Several example of diffraction patterns that do not contain a strong signal is given in figure 3.13. Beside that, there is another type of bad data where the diffraction pattern does not seem to have ellipsoidal symmetry. Typical of those diffraction patterns are is displayed in figure 3.12. Because

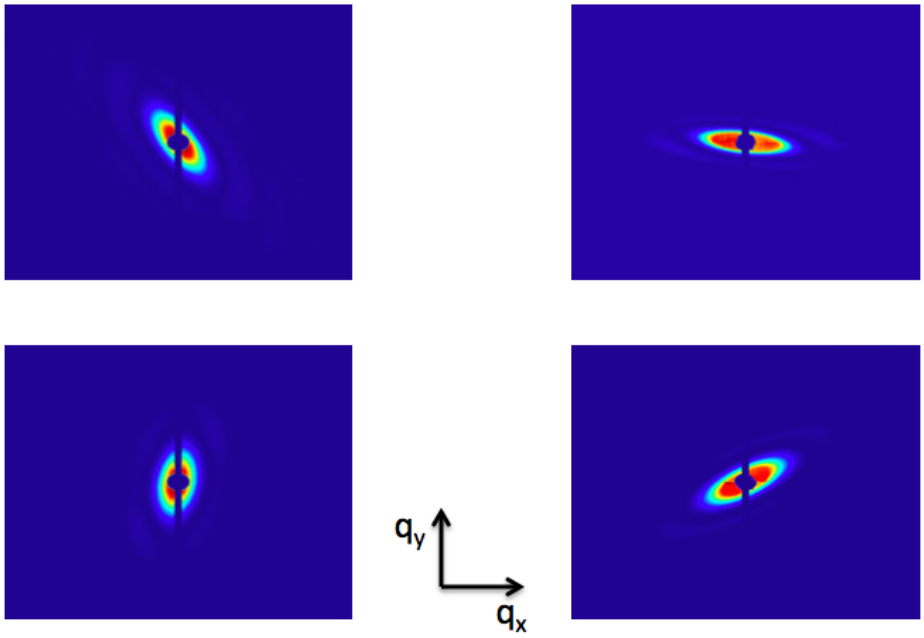


Figure 3.11: Diffraction pattern that are considered as "good"

the nanorice in general is simple structure (close to an ellipsoid), it is expected that its diffraction pattern will be close to ellipsoid diffraction pattern. Thus, the diffraction patterns in figure 3.12 do not come from the nanorice and needs to be excluded for the calculation of $B_l(q, q')$. Given the current stage of algorithm, the algorithm needs to impose azimuthal symmetry. Then such diffraction patterns, which does not show the ellipsoidal behavior, cannot be used to recover the electron density.

After the selection of the good data was obtained, the next step was to obtain the value of each parameter in reciprocal space. To estimate the value of dq (the step of reciprocal distance) for each pixel, following relation was used:

$$dq = \Delta_p/(\lambda Z) \tag{3.4}$$

where dq is reciprocal distance of a pixel in detector, Δ_p is length or size of a pixel in detector, λ is the wavelength used in experiment, Z is the distance from molecule to

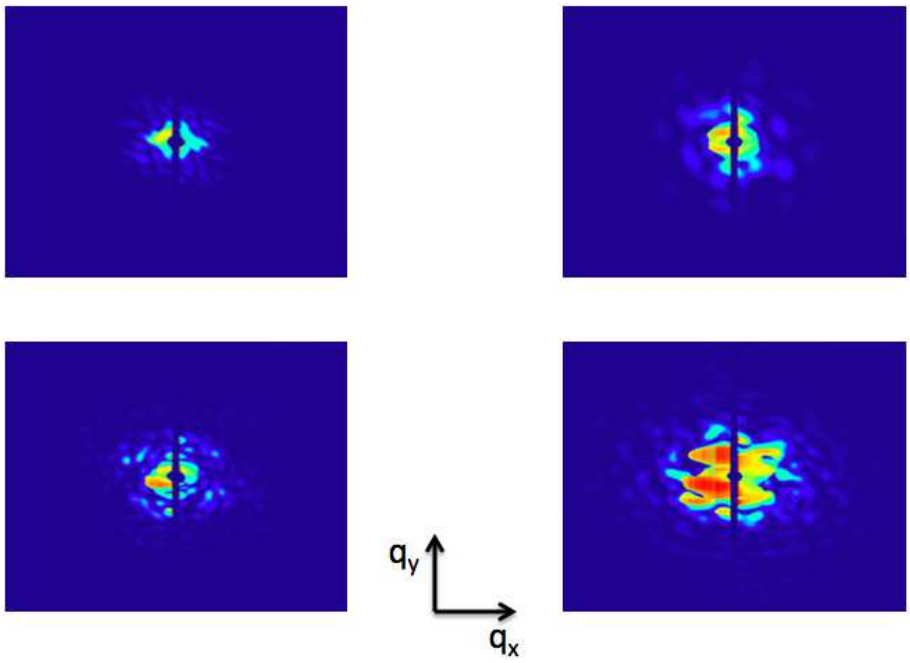


Figure 3.12: Diffraction patterns that are considered as "bad"

detector. All those variables are given by this paper [27] where $\Delta_p = 7.5 \times 10^{-5}$ m, $Z = 0.75$m, and $\lambda = 10.38$ Å

After the quantity dq for each pixel in detector grid was estimated, the next step was to do the interpolation from a Cartesian grid into a polar grid. The first step of interpolation was to specify or determine all points in the polar grid. In this case, Shannon sampling was used for the radial step. The nanorice was estimated to have length 2000 Å, therefore the radial step was $dq = 1/(2D) = 1/(4000) = 2.5 \times 10^{-4}$ Å. In addition to that, the angular step ($d\theta$) was taken as $2\pi/360$. Figure 3.14 shows the arrangement of points in polar grid where $dq = 2.5 \times 10^{-4}$ Åand $d\theta = 2\pi/360$.

The scientific library in matlab was used to specify points in the polar grid. The command in matlab to do the conversion is **pol2cart**. Below is an example of the code to specify a point in the polar grid:

```
//dq is the polar step in reciprocal space
// Nqp is the number of q point
```

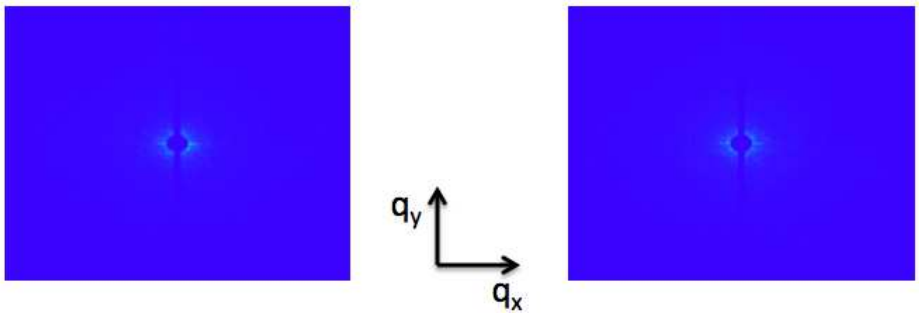


Figure 3.13: Diffraction patterns that does not contain strong scattering

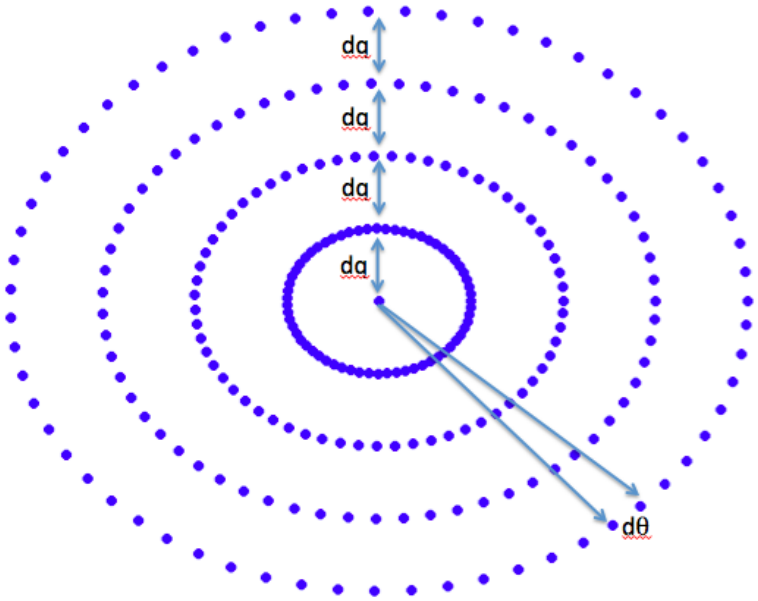


Figure 3.14: The point in polar coordinate

```
dtheta=2*pi/360
qtheta=0:dtheta:2*pi-dtheta
qpolar=0:dq:dq*Nq

// specify polar point in x-y form
[Xp,Yp]=pol2cart(qtheta,qpolar)
```

The output of the code is the quantities Xp and Yp. Those quantities represents the

x-coordinate and y-coordinate of the polar grid. Thus, representation of polar grid in cartesian form can be calculated.

The next step after determining the point in polar coordinate is to interpolate from Cartesian grid (detector grid) to polar coordinates where $B_l(q, q')$ is defined. To get a smooth function of estimation, cubic spline interpolation was used. The cubic spline interpolation divides each interval and approximates the point with a third order polynomial. The polynomial is

$$f_i(x) = a_i + b_i x + c_i x^2 + d_i x^3 \tag{3.5}$$

where subscript i is the index of the interval and (a_i, b_i, c_i, d_i) are the parameters needed to determined from boundary condition. The determination of the parameters used the following boundary condition:

$$\begin{aligned} f_i(x_{i-1}) = y_{i-1} \quad f_i(x_i) = y_i \quad , i = 1, ..., n \\ f_i'(x_i) = f_{i+1}'(x_i) \quad i = 1, ..., n-1 \\ f_i''(x_i) = f_{i+1}''(x_i) \quad i = 1, ..., n-1 \end{aligned} \tag{3.6}$$

where y_i is the known function at x_i, $f_i'(x_i)$ is the first derivative of the function and $f_i''(x_i)$ is the second derivative of the function. Additional two equations are needed to solve the parameters, which are called 'not-a-knot' condition. It imposes conditions of a third derivative to be continous at two end points. Mathematically it is described as:

$$\begin{aligned} f_1'''(x_0) = f_2'''(x_0) \quad i = 1, ..., n-1 \\ f_{n-2}'''(x_n) = f_{n-1}'''(x_n) \quad i = 1, ..., n-1. \end{aligned} \tag{3.7}$$

By applying the condition in equation 3.6 and 3.7, the parameters (a_i, b_i, c_i, d_i) in equation 3.5 can be found. Thus the value of the function in any point can be estimated by equation 3.5. An example of the full derivation of the parameters is given in appendix D.

The conversion of a point in a polar grid from a Cartesian grid requires two dimensional interpolation. The extension from one dimensional to two dimensional is to approximate using two dimensional polynomials. The two dimensional polynomial is

$$f(x,y) = \sum_{i=0}^{3}\sum_{j=0}^{3} a_{ij}x^i y^j. \tag{3.8}$$

Similar to before, the coefficients a_{ij} are determined from boundary conditions in each interval $(i = 1, ..., n)$,

$$f_i(x_i, y_i) = z(x_i, y_i) \quad f_i(x_{i+1}, y_i) = z(x_{i+1}, y_i) \tag{3.9}$$

$$f_i(x_i, y_{i+1}) = z(x_i, y_{i+1}) \quad f_i(x_{i+1}, y_{i+1}) = z(x_{i+1}, y_{i+1}) \tag{3.10}$$

where $z(x_i, y_i)$ is the known function at (x_i, y_i), the first derivative boundary

$$\frac{\partial f_i}{\partial x}\Big|_{x_i,y_i} = \frac{\partial f_{i+1}}{\partial x}\Big|_{x_i,y_i} \qquad \frac{\partial f_i}{\partial x}\Big|_{x_{i+1},y_i} = \frac{\partial f_{i+1}}{\partial x}\Big|_{x_{i+1},y_i} \tag{3.11}$$

$$\frac{\partial f_i}{\partial x}\Big|_{x_i,y_{i+1}} = \frac{\partial f_{i+1}}{\partial x}\Big|_{x_i,y_{i+1}} \qquad \frac{\partial f_i}{\partial x}\Big|_{x_{i+1},y_{i+1}} = \frac{\partial f_{i+1}}{\partial x}\Big|_{x_{i+1},y_{i+1}} \tag{3.12}$$

$$\frac{\partial f_i}{\partial y}\Big|_{x_i,y_i} = \frac{\partial f_{i+1}}{\partial y}\Big|_{x_i,y_i} \qquad \frac{\partial f_i}{\partial y}\Big|_{x_{i+1},y_i} = \frac{\partial f_{i+1}}{\partial y}\Big|_{x_{i+1},y_i} \tag{3.13}$$

$$\frac{\partial f_i}{\partial y}\Big|_{x_i,y_{i+1}} = \frac{\partial f_{i+1}}{\partial y}\Big|_{x_i,y_{i+1}} \qquad \frac{\partial f_i}{\partial y}\Big|_{x_{i+1},y_{i+1}} = \frac{\partial f_{i+1}}{\partial y}\Big|_{x_{i+1},y_{i+1}} \tag{3.14}$$

with the second derivative boundary,

$$\frac{\partial^2 f_i}{\partial x\partial y}\Big|_{x_i,y_i} = \frac{\partial^2 f_{i+1}}{\partial x\partial y}\Big|_{x_i,y_i} \qquad \frac{\partial^2 f_i}{\partial x\partial y}\Big|_{x_{i+1},y_i} = \frac{\partial^2 f_{i+1}}{\partial x\partial y}\Big|_{x_{i+1},y_i} \tag{3.15}$$

$$\frac{\partial^2 f_i}{\partial x\partial y}\Big|_{x_i,y_{i+1}} = \frac{\partial^2 f_{i+1}}{\partial x\partial y}\Big|_{x_i,y_{i+1}} \qquad \frac{\partial^2 f_i}{\partial x\partial y}\Big|_{x_{i+1},y_{i+1}} = \frac{\partial^2 f_{i+1}}{\partial x\partial y}\Big|_{x_{i+1},y_{i+1}} \tag{3.16}$$

$$\tag{3.17}$$

By applying the condition of continuity in the function, the first derivative and the

second derivative, the parameter $a_{i,j}$ can determined in equation . Thus, the value of the function in any point can be estimated by equation 3.1.6.

The derivation to calculate the parameters in the two dimension interpolation can be complicated. However, there are available several scientific libraries to calculate cubic spline interpolation. The library that I used to convert Cartesian grid into polar grid was the matlab libray under the command **interp2**. A snippet of the code to use **interp2** is shown below:

```
//dqc is the reciprocal distance per pixel
// N is the number of pixel in one dimension divided by 2
[Xc,Yc]=meshgrid(-N*dqc:dqc:N*dqc,-N*dqc:dqc:N*dqc)

//dq is the polar step in reciprocal space
// Nqp is the number of q point
dtheta=2*pi/360
qtheta=0:dtheta:2*pi-dtheta
qpolar=0:dq:dq*Nq

// specify the polar point in x-y form
[Xp,Yp]=pol2cart(qtheta,qpolar)

//Do the interpolation
// F is the known value in each cartesian point
// V is the diffraction pattern in polar point
V=interp2(Xc,Yc,F,Xp,Yp,'spline')
```

The output of the code is the quantity V where it holds the value of the interpolation in polar grids. Thus, a set of new polar diffraction patterns can be obtained with the

cubic spline interpolation.

After the diffraction patterns were sampled in polar points, the calculation of the angular pair correlation can be calculated. The pair correlations do the sum over all diffraction patterns and correlate them angularly. The Formula in equation 2.9 was used to calculate C_2 where $I(q, \phi)$ is the diffraction pattern the in polar grid and (q, ϕ) are the points in the polar grid. Subsequent to that, $B_l(q, q')$ was calulated using matrix inversion in equation 2.12. As a result of that, $B_l(q, q')$ from the experimental data of nanorice could be determined.

It is important to note that, the derivation to get $B_l(q, q')$ from equation 2.9 to equation 2.12 use fundamental assumption where all random angles span through all possible angle (equation 2.11). However, there are only 200 good diffraction patterns available from the experimental data. That is why it is important to check the convergence of $B_l(q, q')$ from 200 diffraction patterns.

The matrices $B_l(q, q')$ were calculated from 200 diffraction patterns of nanorice. Subsequent to that, SVD on the $B_l(q, q')$ was performed for each l and its nonzero singular values displayed in figure 3.15

It is shown in the plot that all the matrices $B_l(q, q')$, which correspond to l, have singular values more than $2l + 1$. In other words, the plot shows the behavior more asymmetric than an asymmetric pattern, which is not plausible. The only explanation is that the $B_l(q, q')$ is not in the form of dot product. As explained earlier, if $B_l(q, q')$ is a form of dot product then the SVD on $B_l(q, q')$ will have the number of singular values less than or equal to the number in the asymmetric pattern, which is $2l + 1$. In conclusion, the $B_l(q, q')$ calculated from 200 diffraction pattern of nanorice does not converge to its theoretical value.

Before using the $B_l(q, q')$ for structure determination, checking accurate convergence is a priority. One way to check for convergence is known. Use SVD on the $B_l(q, q')$. If it is found than the number of singular values of $B_l(q, q')$ is higher than those from asymmetric

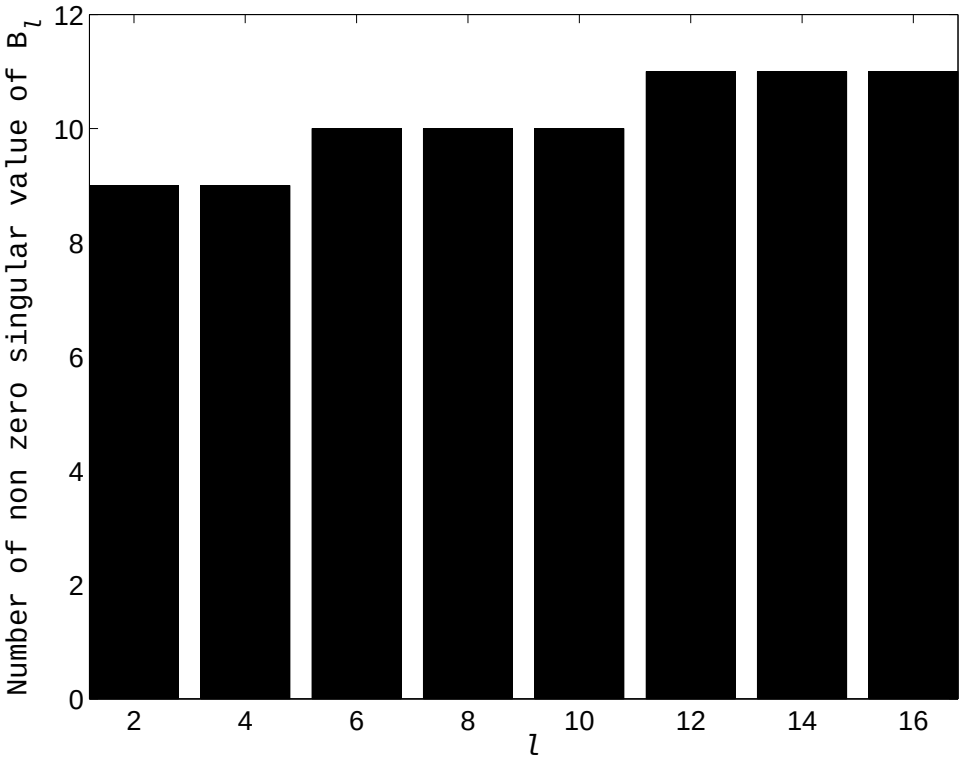


Figure 3.15: The number of nonzero singular value is more than $2l + 1$. The data does not show the convergence of $B_l(q, q')$

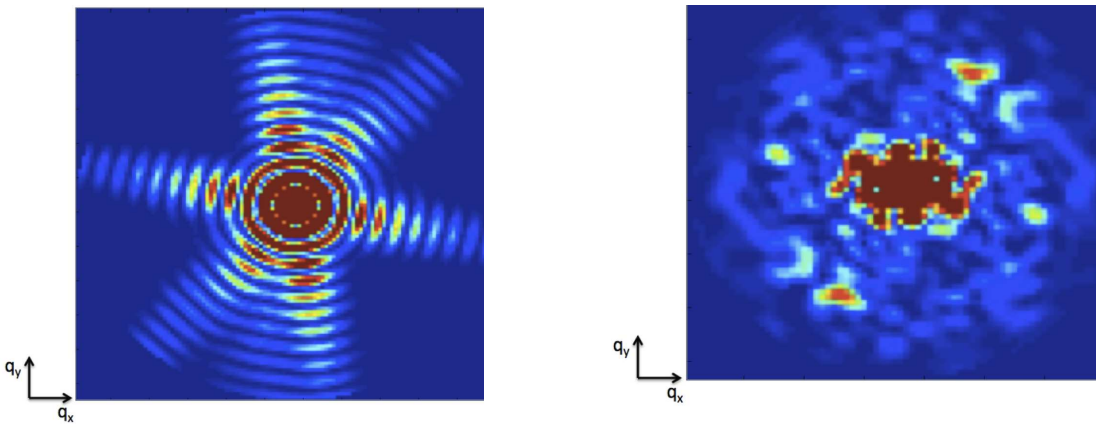
pattern then additional steps to correct the $B_l(q, q')$ are needed.

3.2 Convergence Limit

From eq. 2.9 , the derivation from C_2 to $B_l(q, q')$ uses the fundamental assumption that the collection of random angle spans through all members of rotational group. SO(3) or 3D rotational group have an infinite number of members that are specified by 3 different angles α, β, and γ. The experiment is capable of producing only a finite number of the diffraction patterns. There is a sort of incompatibility between the finite number of the diffraction patterns from experiment and the assumption that gives the same result as all infinite number of members of rotational group. Thus, It is fundamental to find the limit of how the correlations can be used for a finite number of diffraction patterns only.

One can expect at a particular finite number of the diffraction patterns, the calculated $B_l(q, q')$ will converge enough to theoretical $B_l(q, q')$, which is obtained from an infinite number of the diffraction patterns. In order to find the limit, a collection of diffraction patterns is simulated. There are two different structures compared. The first model is PBCV (Paramecium bursaria chlorella)[15]. It is used as a model, which has icosahedral symmetry and the electron density is calculated from the PDB entry(1m4x). The second model is Photoactive yellow protein (PYP) and the electron density is calculated from PDB file (entry 2phy) [16]. PBCV is a structure that has 60 rotational symmetry elements whereas pyp doesn't have rotational symmetry at all. The two different structures that contain different type of symmetry should be able to tell what is the effect of symmetry on the convergence of $B_l(q, q')$. Because $B_l(q, q')$ consists of the expansion of spherical harmonics of the intensity, the value of l correspond to q_{max}, which is directly related to resolution in real space[17]. By plotting for different value of l's, comparison of how convergence of $B_l(q, q')$ affects resolution is studied.

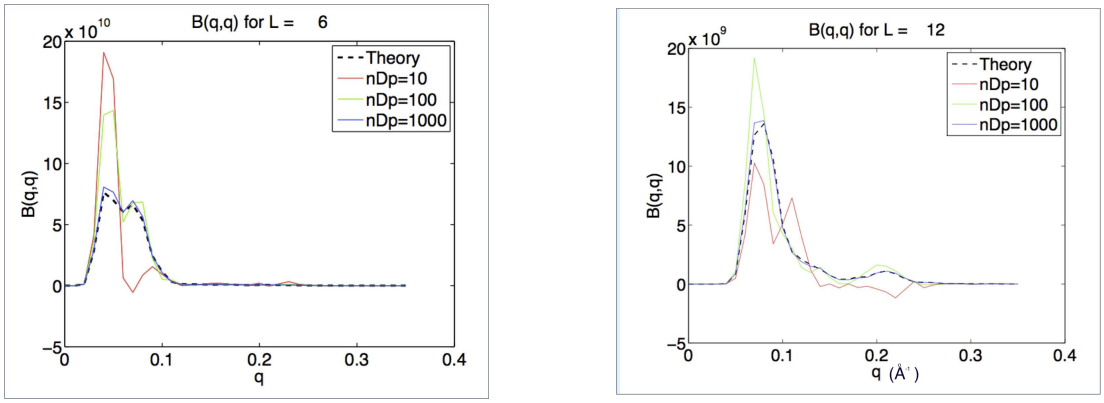
Each diffraction patterns is a 2D slice of intensity in reciprocal space. The randomly-oriented diffraction patterns are simulated by calculating the intensity from the model and taking at random angle a 2D slice of intensity. Typical diffraction patterns from



(a) PBCV (Paramecium bursaria chlorella) (b) PYP (Photoactive yellow protein)

Figure 3.16: A noise free diffraction pattern in random orientation

PBCV and PYP are displayed in figure 3.16. It is easily recognized that the diffraction pattern from PBCV is from a symmetrical object whereas the diffraction pattern from PYP doesn't have an indication of symmetry.



(a) $B_l(q, q')$ for $l = 6$ (b) $B_l(q, q')$ for $l = 12$

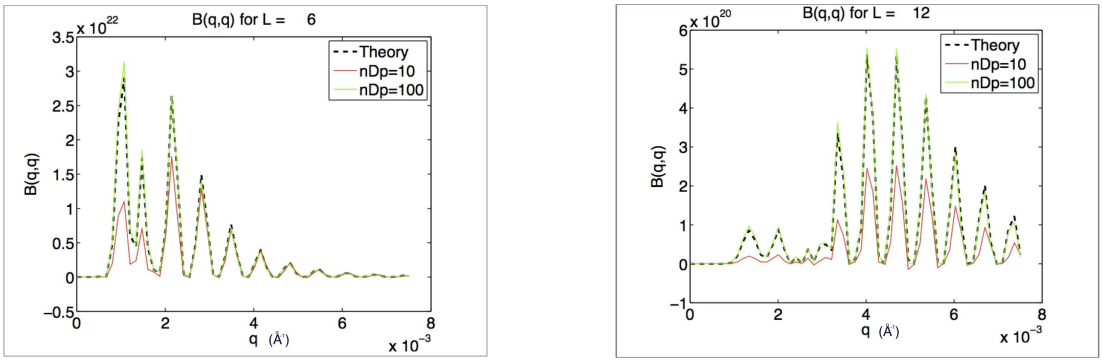
Figure 3.17: Convergence of $B_l(q, q')$ from a set of noise free diffraction patterns of PYP

The convergence of $B_l(q, q')$ can be analyzed by comparing three different sets of simulated diffraction patterns. From figure 3.17, there are curves of $B_l(q, q')$ that are calculated from 10, 100, and 1000 noise-free diffraction patternsiof randm orientation. The curves are represented by solid lines. The dashed lines are the curves of $B_l(q, q')$ calculated from an infinite number of diffraction patterns, as can be done by rhe analytical theory. The

solid lines are expected to converge into the dashed line for a large number of diffraction patterns.

From figure 3.17, the dashed line coincides with the solid line when the number of diffraction pattern reaches about 1000. Two $B_l(q, q')$ curve are calculated, one for $l = 6$ and the other for $l = 12$. Both of them show that the convergence is reached by using only about 1000 diffraction patterns.

Even though the theory explicitly assumes the requirement of infinite number of the diffraction patterns, the simulation shows that only 1000 diffraction patterns are enough to approximate the theoretical value of $B_l(q, q')$. The fact that only about 1000 diffraction patterns are needed indicates the feasibility of the correlation method to be used to recover the electron density.



(a) $B_l(q, q')$ for $l = 6$ (b) $B_l(q, q')$ for $l = 12$

Figure 3.18: The Convergence of $B_l(q, q')$ from a set of noise free diffraction patterns of PBCV

From figure 3.18, the dashed line coincides with the solid line when the number of the diffraction patterns reach 100. Two $B_l(q, q')$ curves are calculated, one for $l = 6$ and $l = 25$. Both of them show the convergence is reached by only 100 diffraction patterns. In contrast to the figure 3.17, the convergence is reached by a significant smaller set of diffraction patterns. The model of figure 3.18 is PBCV, which has 60 different rotational symmetry. It is apparent from the graph that rotational symmetry reduces the total diffraction patterns needed to converge to theoretical value.

The repetition or symmetry of particle is the reason for particle having less independent parameters. In the case of rotational symmetry, the diffraction pattern doesn't change if symmetric particle is rotated with respect to the axis of symmetry. The likelihood of having exact same diffraction pattern by rotation of random angle will increase if the particle has higher rotational symmetry. This explains the decline of the convergence of $B_l(q, q')$ when the particle is PBCV since it has 60 rotational symmetry.

The convergence is an important indication of properly calculated $B_l(q, q')$ from experimental diffraction patterns. In experiment, different sets of diffraction patterns can be collected. By adding a higher number of diffraction patterns, the 2 largest set of diffraction pattern should have a smaller difference because those curves nearly converge to each other. It is expected that the experiment data will have higher number of diffraction patterns to converge compared to simulation. Even though the number of diffraction patterns needed for convergence currently is unknown for the experimental data, the convergence is a necessary condition to be calculated. If convergence is not achieved, more diffraction patterns are needed for that particular structure.

It is possible that the experimental data contain data which are dominated by noise. The inclusion of bad data in the $B_l(q, q')$ calculation will prevent the curve of $B_l(q, q')$ vs q from converging. If the portion of bad data is insignificant, it will have little effect on $B_l(q, q')$ and the convergence will be satisfied. However, whenever the bad data is significant enough in the collection of diffraction patterns, the convergence of $B_l(q, q')$ cannot be achieved. The convergence of $B_l(q, q')$ is an important indication to observe whether bad data is present in the collection of random diffraction patterns. Provided that the convergence of $B_l(q, q')$ is not achieved, the presence of bad data can be one of the reasons further effort to exclude those should be performed.

Assuming bad data is not present but convergence is not achieved, information about the distribution of orientations of diffraction patterns can be deduced. Theoretically, orientations should be random or span through all angles. It is possible that the randomness

of the orientation of the diffraction patterns is not enough to span through uniform angles. Another important point is by observing non-converging $B_l(q, q')$, there is possibility that the diffraction patterns are distributed with non-uniform orientations. It will give clear insight how the experimental processes occur.

This section covers the importance of the convergence of $B_l(q, q')$ that are calculated from the collection of the diffraction patterns. The demonstration of converging $B_l(q, q')$ is performed by using two models namely PBCV and PYP. The feasibility of the correlation method is verified by showing that only 1000 diffraction patterns of asymmetrical molecule converge into theoretical $B_l(q, q')$. Furthermore, the possibility of nonconverging $B_l(q, q')$ from a set of experimental data is discussed. In conclusion, the convergence of $B_l(q, q')$ is vital tool for analyzing experimental data and it is the first step that needs to be confirmed.

Chapter 4

Reconstruction

4.1 2D Case

4.1.1 Polar Fourier Transform

As mentioned in the previous section, several experiments produce diffraction patterns along the azimuthal axis. Then there is only one unknown orientation angle in the collection of diffraction patterns. The independent orientation is rotation with respect to azimuthal axis.

The algorithm explained below will mainly focus on how to get general 2D projected structure from $B_m(q, q')$. The construction of $I_m(q)$ from $B_m(q, q')$ leave phases, which is nonunique [47]. The information about the phases can be gotten by constraining the structure and the intensities as real-positive quantities. That information provides additional information from $B_m(q, q')$ to $I_m(q)$.

Phasing is an algorithm to find the structure from the missing phases of intensities. The way it works is by constraining to any information about the structure. The condition of electron density is positive and real is imposed. In addition to that, several algorithms impose structure to be localized. Phasing is one of established algorithmis that works by

constraining to prior information throughout iterations in both real and reciprocal space.

By definition, $B_m(q, q')$ is described in polar coordinates [46]. To avoid interpolation, polar coordinates are chosen as a basis coordinates for real space and reciprocal space. A consequence is that it excludes the Fast Fourier tranform (FFT) algorithm to be used inside iteration because FFT is defined in Cartesian coordinates. The calculation of a polar Fourier transform is required and needs to be formulated to go back and forth from real and reciprocal space.

Any function can be decomposed into its basis function components. It is sensible to decompose a function into its exponential components because the problem is related to the rotation about an azimuthal angle. The decomposition of the electron density of the molecule is defined by

$$\rho(r, \theta) = \sum_m \rho_m(r) \exp(im\theta) \tag{4.1}$$

where r and θ refer to a coordinate that can be sampled at polar points without interpolation. ρ_m contain only radial dependence and the angular dependence is contained in the complex exponentials.

Intensity is the absolute square of the Fourier transform of the electron density. The established FFT routines cannot be used owing to the fact that ρ is sampled at polar coordinates. It is necessary to obtain a direct relation from the electron density to intensity directly in polar coordinates. The Fourier transform of $\rho(r, \theta)$ is shown below:

$$A(\vec{q}) = \int d^2r \rho(r, \theta) \exp(i\vec{q} \cdot \vec{r}). \tag{4.2}$$

In general, the structure factor is a Fourier transform of the electron density. As stated in equation 4.2, the structure factor is obtained by integrating over all electron densities multiplied by a phase factor. All points are sampled in polar coordinates for both the

electron density $\rho(r,\theta)$ and the structure factor $F(q,\theta_q)$.

One can relate exponential functions to Bessel functionis. The relation is called Jacobi Anger relation, which expresses the exponential function as a sum of Bessel functions [20]. The relation is

$$\begin{aligned} \exp(i\vec{q}\cdot\vec{r}) &= \exp(iqr\cos(\theta_q-\theta_r)) && (4.3)\\ &= \sum_m i^m J_m(qr)\exp(im(\theta_q-\theta_r)). && (4.4)\end{aligned}$$

By substituting the Jacobi-Anger expansion into the Fourier transform of the electron density, a relation between structure factor, electron density and Bessel function is obtained:

$$\begin{aligned} A(\vec{q}) &= \int d^2r\rho(r,\theta)\exp(i\vec{q}\cdot\vec{r}) && (4.5)\\ &= \int d^2r\sum_m \rho(r,\theta)i^m J_m(qr)\exp(im(\theta_q-\theta_r)). && (4.6)\end{aligned}$$

$B_m(q,q')$ can be expressed in terms of an exponential decomposition of the intensity. By substituting equation 4.13 into equation 4.5, the structure factor can be expressed in terms of exponential decomposition. The derivation is shown below:

$$\begin{aligned} A(\vec{q}) &= \int d^2r\sum_m\left[\sum_{m'}\rho_{m'}(r)\exp(im'\theta_r)\right]i^m J_m(qr)\exp(im(\theta_q-\theta_r)) && (4.7)\\ &= \int rdr\sum_{m',m}\rho_{m'}(r)i^m J_m(qr)\exp(im(\theta_q))\int d\theta_r\exp(i(m'-m)\theta_r). && (4.8)\end{aligned}$$

Equation 4.7 involves an infinite integral of the exponential function. That integral is equivalent to the delta function [22], hence equation 4.7 can be simplified become

$$\delta(m'-m)=\frac{1}{2\pi}\int_{\infty}^{-\infty}\exp(ix(m'-m))dx. \quad (4.9)$$

Thus, equation 4.5 can be written as

$$A(\vec{q}) = \int r dr \sum_m \rho_m(r) i^m J_m(qr) 2\pi \exp(im(\theta_q)). \tag{4.10}$$

By decomposing the left hand side of equation 4.10 and equating each component in exponential term, a new relation is found that is

$$\sum_m A_m(q) \exp(im\theta_q) = \sum_m \left[2\pi \int r dr \rho_m(r) i^m J_m(qr) \right] \exp(im\theta_q). \tag{4.11}$$

The important relation between the structure factor decomposition and its exponential components can be obtained:

$$A_m(q) \;=\; 2\pi \int r dr \rho_m(r) i^m J_m(qr). \tag{4.12}$$

In a similar way to the Fourier transform, the inverse Fourier transform is obtained by swaping i to $-i$. It is generally accepted that the electron density is the inverse Fourier transform of the structure factor. By using equation 4.3 and equation 4.9, the derivation is performed in the same way as polar Fourier transform. The following is the derivation:

$$\begin{aligned}
\rho(\vec{r}) &= \int d^2 r A(q, \theta_q) \exp(-i\vec{q} \cdot \vec{r}) \\
&= \int d^2 q \sum_m A(q, \theta_q)(-i)^m J_m(qr) \exp(-im(\theta_q - \theta_r)) \\
&= \int d^2 q \sum_m \left[\sum_{m'} A_{m'}(q) \exp(im'\theta_q) \right] (-i)^m J_m(qr) \exp(-im(\theta_q - \theta_r)) \\
&= \int q dq \sum_{m',m} A_{m'}(q)(-i)^m J_m(qr) \exp(im(\theta_r)) \int d\theta_q \exp(i(m' - m)\theta_q) \\
&= \int q dq \sum_m A_m(q)(-i)^m J_m(qr) \exp(im(\theta_r)) 2\pi \\
\sum_m \rho_m(r) \exp(im\theta_r) &= \sum_m \left[2\pi \int q dq A_m(q)(-i)^m J_m(qr) \right] \exp(im\theta_r).
\end{aligned}$$

Because now the left hand side and the previous right hand side of the equation have the exponential terms, equating each component will lead to a new important equation. The relation between the exponential components of the electron density and the exponential component of the structure factor can be obtained:

$$\rho_m(r) = 2\pi \int q dq A_m(q)(-i)^m J_m(qr). \tag{4.13}$$

Equations 4.12 and 4.13 are the foundation to do Fourier transform in polar coordinates directly without involving the cartesian space. After obtaining the exponential components of the structure factor or the electron density, the summation with respect to its basis function is performed to get $\rho(\vec{r})$ or $A(\vec{q})$ as shown

$$\begin{aligned} \rho(\vec{r}) &= \sum_m \rho_m(r) \exp(im\theta) \\ A(\vec{q}) &= \sum_m A_m(q) \exp(im\theta). \end{aligned} \tag{4.14}$$

4.1.2 Angular Correlation Constraint

As explained earlier, the correlation methods are needed to calculate $B_m(q, q')$ from experiment. Information of intensity can be obtained from $B_m(q, q')$ and the other constraints. A phasing algorithm can be modified so that the constraint is on $B_m(q, q')$ instead of intensity[43]. The step is explained as follows:

1. Start initial guess of $\rho(\vec{r})$
2. Calculate $\rho_m(q)$ from $\rho(\vec{r})$
3. Use equation 4.12 to calculate $A_m(q)$
4. Calculate $I_m(q)$ from $A_m(q)$ and keep information of phase.
   Start from calculating $A(\vec{q}) = \sum_m A_m(\vec{q}) \exp(im\theta)$ then $I(\vec{q}) = |A(\vec{q})|^2$.

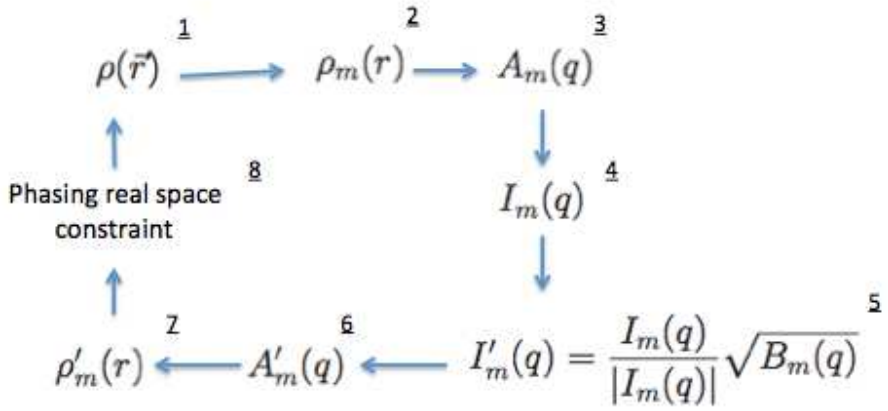


Figure 4.1: Full cycle of phasing algorithm with $B_m(q,q)$ as constraint

Final step is $I_m(q) = \int I(\vec{q}) \exp(-im\theta) d\theta$

5. Project $I_m(q)$ to satisfy $B_m(q,q)$ constraint

6. Use phase from step 4 to obtain $A'_m(q)$

7. Calculate $\rho_m(q)$ from equation 4.13

8. Use HIO, ER, and shrinkwrap to constrain $\rho(\vec{r})$ and the cycle is repeated

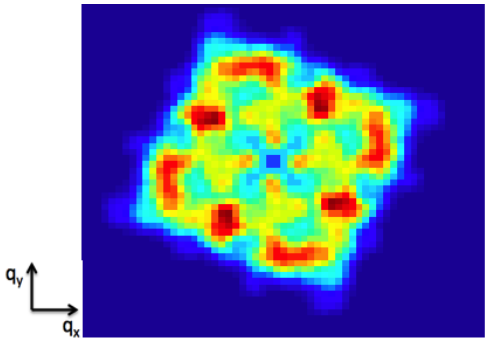


Figure 4.2: Electron density of K channel protein is used as a model to calculate $B_m(q,q)$

Figure 4.2 is model that is used simulate $B_m(q,q)$. After applying phasing constraint on $B_m(q,q)$, the reconstruction is shown on figure 4.3.

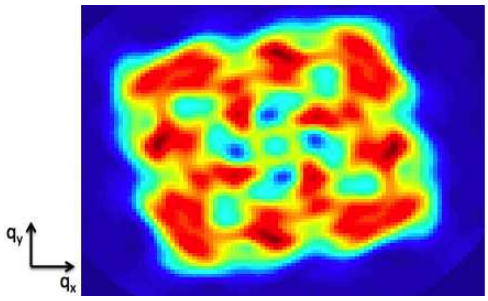


Figure 4.3: Reconstruction of electron density by only constraining to diagonal value of $B_m(q,q)$

Important to note that in this reconstruction only information on diagonal values of $B_m(q,q)$ is used. Another important quantity is R-factor, R-factor between $B_m(q,q')$ model and its reconstruction is 0.15, which is defined as

$$R_{factor} = \frac{\sum ||B_m(q,q)| - |B_m(q,q)_{exp}||}{\sum |B_m(q,q)_{exp}|}. \tag{4.15}$$

Judging from figure 4.3 and R-factor, the reconstruction is reasonable enough even though only diagonal values are used as constraints. The discrepancy between the model and reconstruction could be attributed to the fact that only diagonal values of $B_m(q,q)$ are used.

4.2 Triple Correlation

Apart from $B_l(q,q')$, triple correlations, which can be calculated from squaring one term in the pair correlations, also can be used to reconstruct the electron density. Since triple correlations are more complicated than $B_l(q,q')$, only the case which has azimuthal property is considered in this section.

In this section, a method to recover azimuthal electron density from an ensemble of random angle diffraction patterns is explained. The method is developed to recover the electron density of nanorice. Moreover, experimental diffraction patterns are readily

available and can be downloaded from cxidb.org. [27]

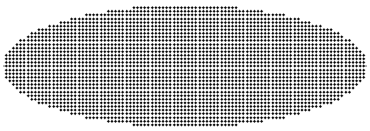


Figure 4.4: 3D ellipsoidal cartesian grid is used as model

The object under study is nanorice, which has an ellipsoidal shape. For an initial model of nanorice we assumed an ellipsoid on a 3D cartesian grid in real space. We took the electron density of the nanorice particle to be 1 inside and zero outside the ellipsoid. This model is shown on Figure 4.4 . Subsequently random angle diffraction patterns can be simulated.

The first step of simulation is by calculating the structure factor. The structure factor is calculated in reciprocal space using the Fourier transform of the model. The calculation of structure factor and the intensity is

$$A(\hat{q}) = \sum_j \rho(\hat{r}_j) \exp(2\pi \hat{q}.\hat{r}_j) \tag{4.16}$$

$$I(\hat{q}) = |A(\hat{q})|^2. \tag{4.17}$$

After the intensity is obtained, its spherical harmonics expansion are calculated by integrating the intensity times spherical harmonics over all angles on surface area. Mathematically, the calculation is expressed as

$$I_{lm}(q) = \int I(q, \theta, \phi) Y^*_{lm}(\theta, \phi) d\Omega. \tag{4.18}$$

Now, random rotation of I_{lm} is performed by multiplying the I_{lm} by rotation matrix. The rotation matrix that has spherical harmonics as their basis functions is Wigner D-matrix. The result of multiplication of the I_{lm} by Wigner D-matrix is new I_{lm} with rotated axes

described as

$$I'_{lm}(q) \quad = \quad D^{l}_{m,m'}(\alpha,\beta,\gamma)I_{lm'}(\theta,\phi). \tag{4.19}$$

Finally, a diffraction pattern is calculated by slicing the diffraction volume through a plane q_z=0:

$$I'(q,\theta=\frac{\pi}{2},\phi) \quad = \quad \sum_{lm} I'_{lm}(q)Y_{lm}(\theta=\frac{\pi}{2},\phi). \tag{4.20}$$

This represents the diffraction patterns from random orientations of an nanorice particle. Typical such diffraction patterns are shown in Fig. 4.5

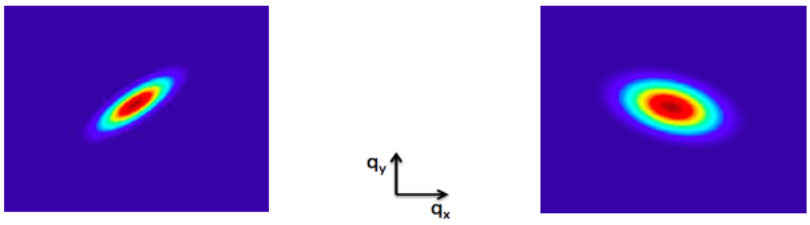


Figure 4.5: Diffraction patterns of nanorice in random orientation

Having thus simulated the random orientations diffraction patterns, our next step was to demonstrate it is possible to reconstruct our model of nanorice from those patterns. To do this we have to calculate $B_l(q,q)$ and $T_l(q,q)$ from the simulated diffraction patterns.

The coefficients of a spherical harmonic expansion (I_{lm}) of the diffraction volume clearly depends on the orientation of the diffraction volume relative to the chosen z-axis. Two such 3D intensity distributions are displayed on Fig. 4.6 and Fig, 4.7. By choosing z-axis at the center of azimuthal symmetry, we eliminate the other components of I_{lm} except m=0.

This suggests some arbitrariness in the reconstruction of the diffraction volume from the measured B_l and T_l coefficients, which are orientation-independent quantities. The

quantities B_l and T_l depend on the angular momentum quantum number l but not on the azimuthal quantum number m. Yet, in general the spherical harmonic expansion coefficients $I_{lm}(q)$ depend on both sets of quantum numbers. However there is one orientation when the I_{lm} coefficients themselves only depend on l, and that is when a major axis of the ellipsoid representing the nanorice is coincident with the z-axis. Under these conditions the particle, and also the diffraction volume has azimuthal symmetry about the z-axis, and can be characterized exactly by $m = 0$ for all l.

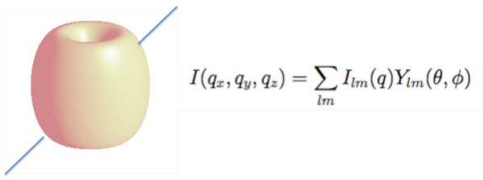


Figure 4.6: Expansion in spherical harmonics with respect to an arbitrary axis

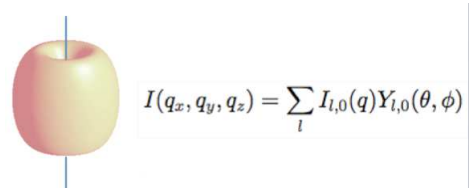


Figure 4.7: Expansion in spherical harmonics with respect to the z-axis

We are not trying to reconstruct the particle in any particular orientation. An orientation with the major axis of the ellipsoid along the z-axis is just as good as any other. We can choose this orientation by assuming that only the $m = 0$ components of the $I_{lm}(q)$'s exist. At this point these coefficients depend only on l, since we assume we know the value of m, and we can write

$$|I_{l0}(q)| = \sqrt{B_l(q,q)}. \tag{4.21}$$

Also, as it is the angular average of the diffraction intensity on a resolution shell of radius q, it is a real quantity. Consequently the only ambiguity in $I_{l0}(q)$ is in its sign.

We can determine this sign from the triple correlations, since nanorice has azimuthal symmetry. In this case, $m_1 = m_2 = 0$, and the Legendre transformation of the triple correlation reduces to

$$T_l(q,q) = \sum_{l_1 l_2} I_{l0}(q) I_{l_1 0}(q) I_{l_2 0}(q) G(l0; l_1 0; l_2 0) \tag{4.22}$$

where G is a Gaunt coefficient [28].

Since an ellipsoid has azimuthal symmetry about a particular axis, we can choose that particular axis as z-axis, thus eliminating any other components of I_{lm} except $m = 0$. $|I_{l,0}|$ can be obtained directly from B_l via

$$|I_{l,0}(q)| = \sqrt{B_l(q)}. \tag{4.23}$$

The only unknown here is sign of $I_{l,0}$. The sign can be determined by fitting all possible signs of $I_{l,0}$ to the "experimental" triple correlations in (4.22).

After obtaining the signs of the $I_{l.0}$, the diffraction volume can be calculated from

$$I(\hat{q}) = \sum_{l,0} I_{l,0} Y_{l,0}(\theta, \phi). \tag{4.24}$$

To test the method, 200 random angle diffraction patterns were simulated. Typical diffraction patterns are shown in figure 4.5. After simulating the diffraction patterns, they were used as the input to calculate C_2, C_3, $T_l(q)$ and $B_l(q,q)$. For the reconstruction, the parameter $l_{max} = 16$ was used as a cut off of the maximum value for l.

As explained above, by knowing $B_l(q)$ and $T_l(q)$, the diffraction volume can be found after constraining their spherical harmonic expansion to be nonzero only when $m = 0$. An iterative phasing algorithm [29, 30] applied to this diffraction volume could then recover the electron density. The reconstructed electron density after phasing is displayed in

figure 4.8. Figure 4.8 shows that the method can reconstruct the original ellipsoid model from a set of random angle diffraction patterns.

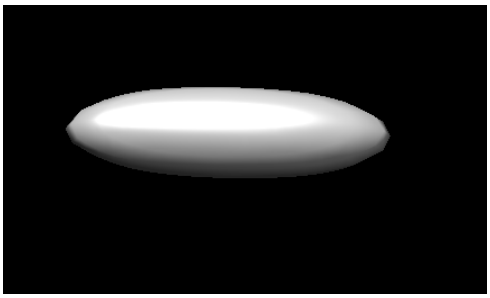


Figure 4.8: Reconstructed electron density after phasing

Another test of this method was by calculating R_{split}. A set of diffraction patterns were simulated and they were split into two sets of diffraction patterns (each set has 200 patterns). By having 2 sets, now there were 2 different quanities for C_2, C_3, $B_l(q)$, and $T_l(q)$. Each quantities was used to calculate two different diffraction volumes. After getting two different diffraction volumes, formula for R_{split} was used as follow:

$$R_{split}(q) = \frac{1}{2^{1/2}} \frac{\sum_{|q|} |I_1 - I_2|}{\frac{1}{2}\sum_{|q|}(I_1 + I_2)} \tag{4.25}$$

where the summation is performed over all pointis in the shell surface with the same value of q.

The quantity q_{max} can be estimated using

$$q_{max} = \frac{l_{max}}{2\pi R} \tag{4.26}$$

where $l_{max} = 16$ and $R = 25$ Å. By using equation 4.26, it is expected that the q_{max} is accurate until 0.1 Å^{-1}. In figure 4.9, it is shown that the R_{split} is reasonable enough when q_{max} is less than 0.1 Å^{-1} and R_{split} goes higher after $q_{qmax} = 0.1$ Å^{-1}.

Beside R_{split}, Fourier shell correlation (FSC) was used to characterize the reconstruc-

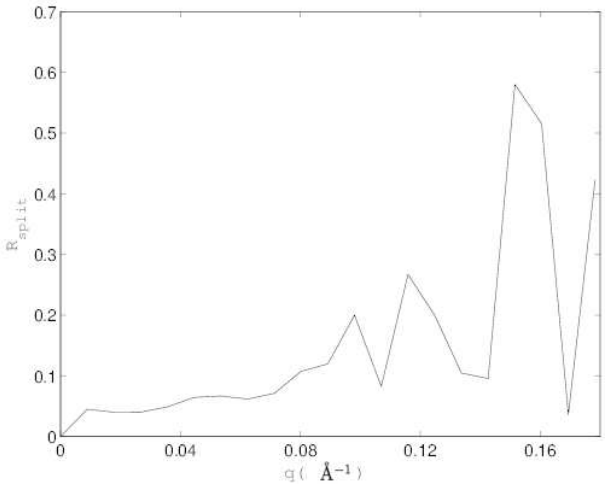


Figure 4.9: Plot of R_{split} vs q

tion. As explained above, R_{split} does the comparison of two diffraction volumes before they are used in phasing. In constrast to R_{split}, FSC includes the comparison of the reconstruction after phasing. Thus, the two diffraction volumes from earlier calculation were used as the inputs of the phasing and the outputs of it were the two different electron densities.

The electron density from a phasing algorithm has an arbitrary center. Before calculating the FSC, the centering was performed by finding the location of a box, which has largest value of the electron density. After getting the location of the box, the center of the box was taken as the center of the electron density. Subsequent to that, two structure factors were calculated by performing a Fourier transform of the electron density. Following that, FSC can be calculated using the following formula

$$FSC(q) = \frac{\sum_{|q|} F_1(q)F_2(q)^*}{\sqrt{\sum_{|q|} |F_1(q)|^2}\sqrt{\sum_{|q|} |F_2(q)|^2}} \tag{4.27}$$

where the summation is performed over all points in the shell surface with the same value of q.

Figure 4.10 shows the plot of the modulus of FSC vs q. The FSC goes down as q goes

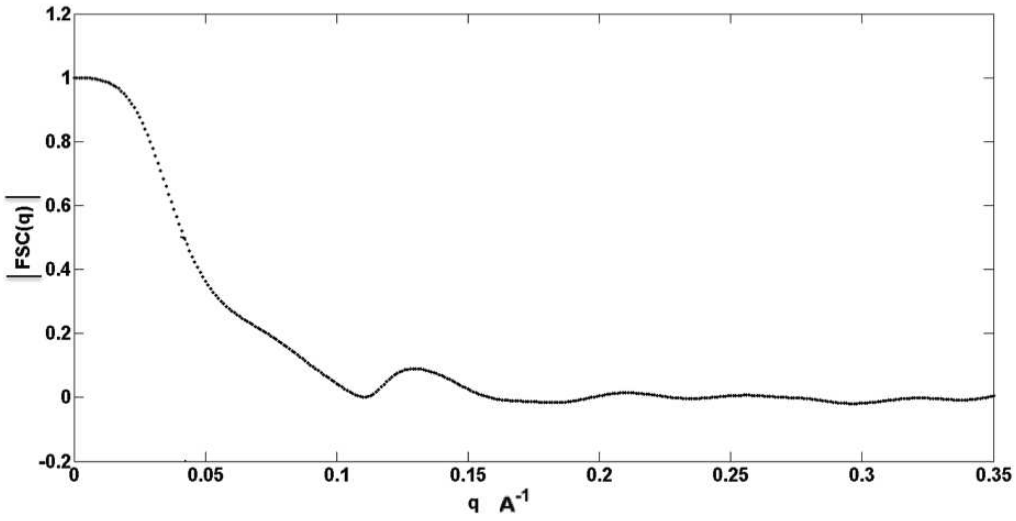


Figure 4.10: Plot of modulus of FSC vs q

higher. The graph shows that even though $q_{max} = 0.1$ Å^{-1}, the actual accuracy for q_{max} is lower than that. If 0.5 is taken as the limit of acceptable value of FSC then the actual q_{max} is around 0.05 Å^{-1}. Many factor can contribute to the calculation of FSC such as the total number of the diffraction patterns, the phasing algorithm, and the process of centering the electron density.

In this simulation, the diameter of object is 50 Å. From equation 4.26, theoritically the expected resolution is $q = 0.1$ Å^{-1} or 10 Å. The reason of the resolution is low because only $l_{lmax} = 16$ is used. The result of R_{split} give similar resolution compare to the theoritical value. However, the result of FSC is worse than the expected value. The reason of that because the calculation of FSC involves the centering the object and phasing algorithm. Any error from those calculation will propagate to the calculation of FSC. Another important point is only 200 diffraction patterns are used in the simulation. As mentioned in the previous section, it is possible that 200 diffraction patterns doesn't give unique convergent for higher resolution therefore the error will propagate to the calculation of FSC for higher resolution.

4.3 Positivity Constraint

4.3.1 Matrix Quantity

This section particularly explains the reconstruction of the 3D electron density from the $B_l(q, q')$. The essential treatmeant of this method is to convert all quantities into a matrix or vector.

The first step is to find the estimate of $I_{lm}(q)$. The singular value decomposition (SVD) of the $B_l(q, q')$ can be used to get an estimate of $I_{lm}(q)$. From equation 4.30, SVD on the $B_l(q, q')$ for particular l is performed and the nonuniqueness of $I_{lm}(q)$ originates from unitary matrix O. The derivation is

$$B_l(q, q') = U D U^t \tag{4.28}$$

$$B_l(q, q') = U\sqrt{D}O^\dagger O\sqrt{D}U^t \tag{4.29}$$

$$I_{lm}(q) = (U\sqrt{D})O^\dagger. \tag{4.30}$$

One can shows that the $I_{lm}(q)$ on equation 4.30 is already in matrix form. For the reason of clarity in the indices, a new matrix G are defined in equation 4.31. The matrix G is the first estimate of $I_{lm}(q)$ and the actual solution of $I_{lm}(q)$ has dependence on $O^\dagger$. The rows correspond to the q coordinate and the columns correspond m value, or the singular values in the symmetric case. For the asymmetric case, there will be $2l + 1$ columns in matrix G whereas for 4-fold symmetry the number of column for each l is shown in figure 3.5. Additionally, the matrix G is in a form of multiplication between U, which is eigenvector of the $B_l(q, q')$, and $\sqrt{D}$, which is diagonal matrix consistng of the singular values of the $B_l(q, q')$. In general:

$$G = U\sqrt{D} \tag{4.31}$$

$$\begin{pmatrix} I_{l(-l)}(q_1) & \dots & I_{l(l)}(q_1) \\ I_{l(-l)}(q_2) & \dots & I_{l(l)}(q_2) \\ I_{l(-l)}(q_3) & \dots & I_{l(l)}(q_3) \\ \vdots & \vdots & \vdots \\ I_{l(-l)}(q_N) & \dots & I_{l(l)}(q_N) \end{pmatrix} =$$

$$\begin{pmatrix} G^l_{1(-l)} & G^l_{1(-l+1)} & \dots & G^l_{1(l)} \\ G^l_{2(-l)} & G^l_{2(-l+1)} & \dots & G^l_{2(l)} \\ \vdots & \vdots & \vdots & \vdots \\ G^l_{N(-l)} & G^l_{N(-l+1)} & \dots & G^l_{N(l)} \end{pmatrix} \begin{pmatrix} O^l_{1(-l)} & O^l_{1(-l+1)} & \dots & O^l_{1(l)} \\ O^l_{2(-l)} & O^l_{2(-l+1)} & \dots & O^l_{2(l)} \\ \vdots & \vdots & \vdots & \dots \\ O^l_{(l)(-l)} & O^l_{(l)(-l+1)} & \dots & O^l_{(l)(l)} \end{pmatrix} \tag{4.32}$$

The Full $I_{lm}(q)$ matrix is shown in equation 4.32. The rows correspond to the q coordinate and columns correspond to m values. The matrix in equation 4.32 is an example of how multiplication occur on the matrix G and $O^l_{mm'}$. Depending on the symmetry, in general the matrix O^l_{mm} has the dimension $2l+1$ by $2l+1$. The goal of this method is to find the matrix $O^l_{mm'}$ by any constraint other than $B_l(q,q')$

In order to use the positivity constraint, the relation between $I(\vec{q})$, $B_l(q,q')$, and $O^l_{mm'}$ is needed. $I(\vec{q})$ can be found by substituting $I_{lm}(q)$ to its spherical harmonics expansion. Shown in equation 4.35, $I(\vec{q})$ can be related to the matrix G, $O^l_{mm'}$, and $Y_{lm}(\Omega)$.

$$I_{lm}(q_1) = \sum_{m'} G^l_{1(m')} O^l_{i(m)} \tag{4.33}$$

$$I(q_1, \Omega_1) = \sum_{lm} I_{lm}(q_1) Y_{lm}(\Omega_1) \tag{4.34}$$

$$I(q_1, \Omega_1) = \sum_{lmm'} G^l_{1(m')} O^l_{m'(m)} Y_{lm}(\Omega_1) \tag{4.35}$$

It is important to note that equation 4.35 is a form of a linear equation. The property of many linear equations is it can be separated between known and unknown quantity. In this case the matrices G and $Y_{lm}(\Omega)$ are known whereas matrix $O^l_{mm'}$ is an unknown quantity. The separation is constructed by creating matrix consisting of several linear equations. Equation 4.36 is the matrix constructed from linear equation of $I(\vec{q})$ at different q point.

$$\begin{pmatrix} I(q_1) \\ I(q_2) \\ \dots \\ I(q_n) \end{pmatrix} = \tag{4.36}$$

$$\begin{pmatrix} G^0_{1(0)}Y_{00} & G^2_{1(-2)}Y_{2(-2)} & \dots & G^2_{1(2)}Y_{2(2)} & G^4_{1(-4)}Y_{4(-4)} & \dots & G^4_{1(4)}Y_{(4)(-4)} \\ G^0_{1(0)}Y_{00} & G^2_{1(-2)}Y_{2(-2)} & \dots & G^2_{1(2)}Y_{2(2)} & G^4_{1(-4)}Y_{4(-4)} & \dots & G^4_{1(4)}Y_{(4)(-4)} \\ \dots & \dots & \dots & \dots & & & \\ G^0_{N(0)}Y_{00} & G^2_{N(-2)}Y_{2(-2)} & \dots & G^2_{N(2)}Y_{2(2)} & G^4_{N(-4)}Y_{4(-4)} & \dots & G^4_{1(4)}Y_{(4)(-4)} \end{pmatrix} \begin{pmatrix} O^0_{1(0)} \\ O^2_{(-2)(-2)} \\ \vdots \\ O^2_{(2)(2)} \\ O^4_{(4)(-4)} \\ \vdots \\ O^4_{(4)(4)} \\ \vdots \end{pmatrix}$$

To be more concise, a new definition of the matrix C and the vector V is made. The matrix C is multiplication between G and Y_{lm}. The vector V is an one dimensional vector which consist of element or unitary matrix $O^l_{mm'}$.

$$C = \begin{pmatrix} G^0_{1(0)}Y_{00} & G^2_{1(-2)}Y_{2(-2)} & \dots & G^2_{1(2)}Y_{2(2)} & G^4_{1(-4)}Y_{4(-4)} & \dots & G^4_{1(4)}Y_{(4)(-4)} \\ G^0_{1(0)}Y_{00} & G^2_{1(-2)}Y_{2(-2)} & \dots & G^2_{1(2)}Y_{2(2)} & G^4_{1(-4)}Y_{4(-4)} & \dots & G^4_{1(4)}Y_{(4)(-4)} \\ \dots & \dots & \dots & \dots & & & \\ G^0_{N(0)}Y_{00} & G^2_{N(-2)}Y_{2(-2)} & \dots & G^2_{N(2)}Y_{2(2)} & G^4_{N(-4)}Y_{4(-4)} & \dots & G^4_{1(4)}Y_{(4)(-4)} \end{pmatrix} \tag{4.37}$$

$$\vec{V} = \begin{pmatrix} O^0_{1(0)} \\ O^2_{(-2)(-2)} \\ \vdots \\ O^2_{(2)(2)} \\ O^4_{(4)(-4)} \\ \vdots \\ O^4_{(4)(4)} \\ \vdots \end{pmatrix} \tag{4.38}$$

Equation 4.39 is a matrix relation which relates the intensity and $B_l(q,q')$. The information about $B_l(q,q')$ is retained inside matrix C that comes from SVD of $B_l(q,q')$. The vector V is giving the nonuniqueness of intensity from $B_l(q,q')$ data since its elements consist of elements of a unitary matrix $O^l_{mm'}$.

Important to note that the relation, which is based on equation 4.39, can be used to check whether the intensity satisfies $B_l(q,q')$ constraint or not. If the data of intensity is available, then by taking the inverse matrix C, which is multiplied by the intensity, a new vector V can be calculated. If the new vector V consist of element of a unitary matrix then that intensity satisfies the $B_l(q,q')$ constraint. However if elements of the new vector V doesn't have the property of unitary matrix then that intensity doesn't satisfy

the $B_l(q, q')$ constraint.

$$I(\vec{q}, \Omega) = CV \tag{4.39}$$

4.3.2 Optimization

As mentioned in previous section, the intensity is always positive because it is the absolute value of the amplitude. This fact can be used to limit the range of solutions and resolve the nonuniqueness of the unitary matrix $O^l_{mm'}$. The optimization can be used to constrain the intensity to be positive and at the same time satisfy requirement $B_l(q, q')$.

There is an optimization algorithm that is suitable for constraining to positive values and which satisfy the objective function at the same time, which is called active set, its definition is

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad Ax \geq b \end{aligned} \tag{4.40}$$

According to equation 4.39, variables in equation 4.41 need to be adjusted. In this case, $b = 0$, $x = V$, and $A = C$ to satisfy the positivity constraint.

From equation 4.30, as long as $O^l_{mm'}$ is unitary matrix then $B_l(q, q')$ constraint is satisfied. Based on that requirement, the objective function $f(x)$ in equation 4.41 is such that the matrix $O^l_{mm'}$ is unitary.

Mathematically, unitary matrix is

$$O^l_{mm'}(O^l_{mm'})^\dagger = 1. \tag{4.41}$$

By defining new quantity,

$$N^l_{nn'} = \sum_m O^l_{nm}(O^l_{n'm})^\dagger \tag{4.42}$$

$$\begin{pmatrix} N^l_{(-l)(-l)} & N^l_{(-l)(-l+1)} & \cdots & N^l_{(-l)(l)} \\ N^l_{(-l+1)(-l)} & N^l_{(-l+1)(-l+1)} & \cdots & N^l_{(-l+1)(l)} \\ \vdots & \vdots & \vdots & \\ N^l_{(l)(-l)} & N^l_{(l)(-l+1)} & \cdots & N^l_{(l)(l)} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

If matrix $N^l_{nn'}$ is identity matrix then following is satisfied

$$\sum_{n,n',l}(N^l_{nn'} - \delta_{n,n'})^2 = 0 \tag{4.43}$$

where $\delta_{n,n'}$ is Kronecker delta

Equation 4.44 can be used as objective function. The objective function here is to ensure $B_l(q,q')$ is satisfied or in other words matrix $O^l_{mm'}$ is unitary. The matrix $O^l_{mm'}$ is unitary if $N^l_{nn'}$ is identity matrix based on equation 4.43. As a consequence of that, if equation 4.44 is satisfied then $B_l(q,q')$ constraint is satisfied as well. The definition of the objective function written in full way that is

$$\text{minimize} \quad \sum_{n,n',l}(N^l_{nn'} - \delta_{n,n'})^2 \tag{4.44}$$

$$\text{where} \quad N^l_{nn'} = \sum_m O^l_{nm}(O^l_{n'm})^\dagger \tag{4.45}$$

$$\text{subject to} \quad I(\vec{q}) = CV \geq 0 \tag{4.46}$$

the built in function in matlab is used to perform optimization with the active set algorithm. In matlab command, active set is in under command *fmincon*.

The simulation was done by calculating $B_l(q,q')$ of the K channel protein. After $B_l(q,q')$ of the K channel protein was calculated then optimization based on equation

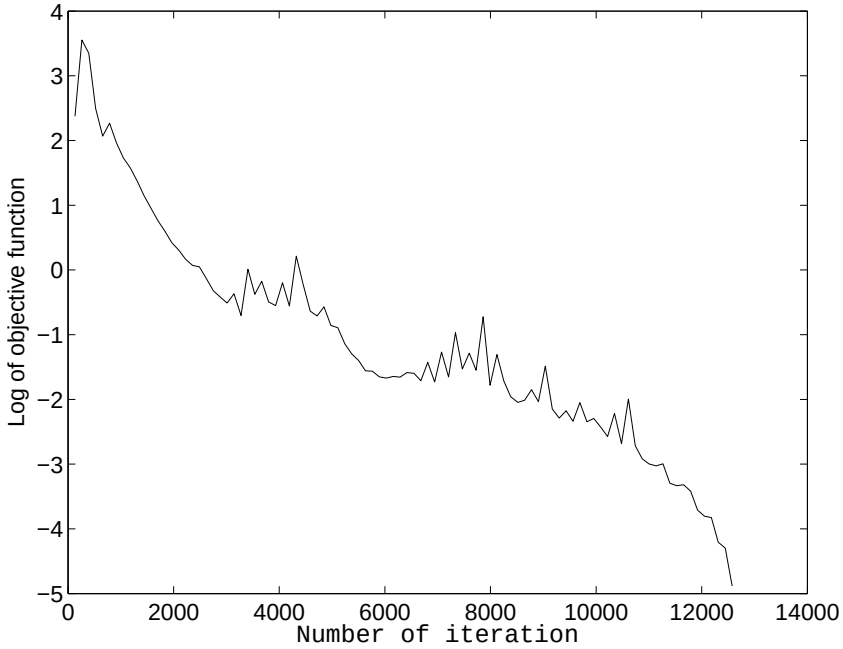


Figure 4.11: Log of objective function vs number of iteration

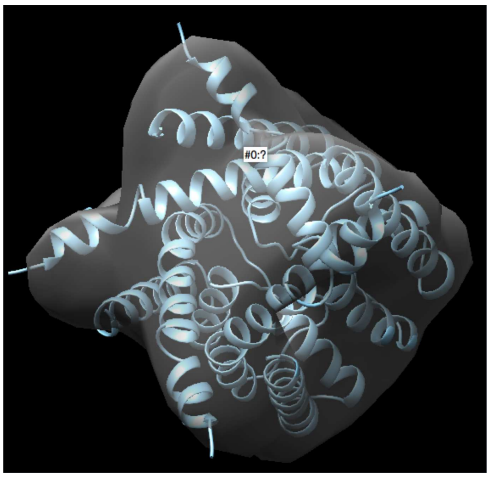


Figure 4.12: reconstruction of electron density

4.46 was used. Graph on figure 4.11 is the plot of *log* of the objective function vs number of iteration. It is obvious from graph that by the end of iteration the objective function is reaching 10^{-6}, which is small enough or approaching zero. In other words, by objective function is zero then requirement matrix that $O^l_{mm'}$ is unitary is satisfied.

After the intensity was reconstructed, the electron density was obtained by using

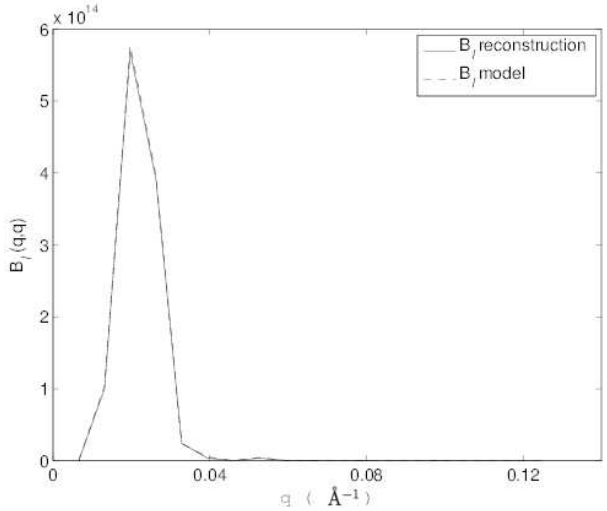
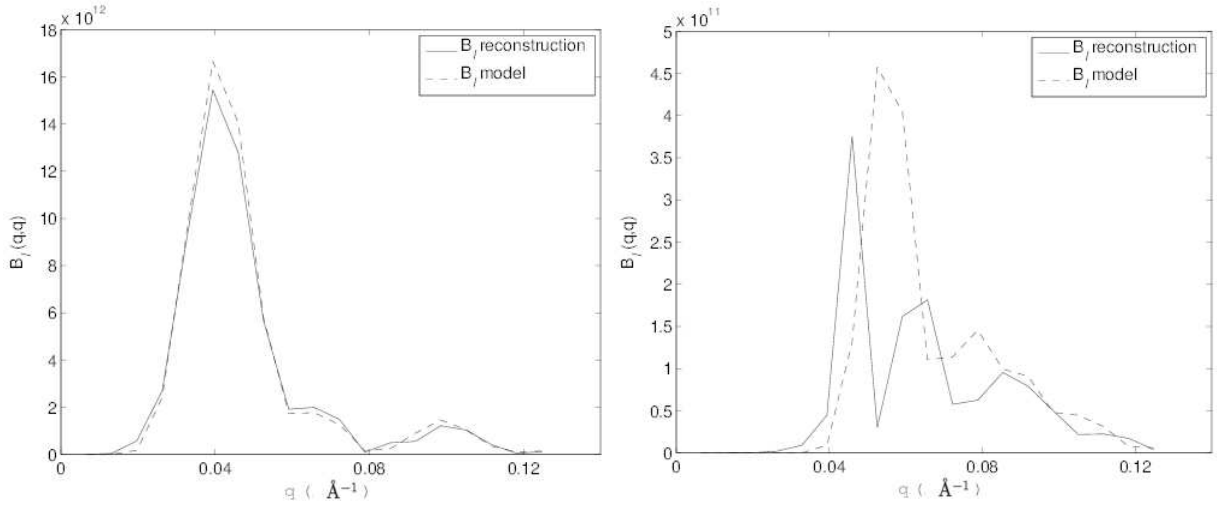


Figure 4.13: Validation model and its reconstruction



(a) $l = 10$ (b) $l = 16$

Figure 4.14: Validation model and its reconstruction

charge flipping algorithm. Figure 4.12 are electron density after phasing algorithm. It has 4-fold symmetry and it enclose the original model. To test how valid the reconstruction is, $B_l(q, q')$ is compared between model and reconstruction. It is shown in graph on figure 4.13 and 4.14, that reconstruction can recover $B_l(q, q')$ for $l = 2$ and $l = 10$. However from $l = 16$, $B_l(q, q')$ begin to deviate from original model. Currently that is the limit of this method since the method only considers positivity constraint. There is other constraint

that is not considered namely real space constraint on electron density. The explanation of real space constraint is given in chapter 5

Chapter 5

Conclusion and Outlook

Section 4.1 shows the reconstruction of projected electron density using $B_m(q, q')$ as a constraint. It is shown that by combining a phasing algorithm with the constraint on $B_m(q, q')$ the electron density converges into the original model. Another important treatmant in the method is to use a Fourier transform in polar coordinates. The definition of $B_m(q, q')$ is described in polar coordinate, then the loss of information due to the interpolation is minimal throughout the iterations in phasing. For that reason, a new phasing algorithm in terms of polar coordinates is developed in section 4.1.

The important constraint that is shown in section 4.1 is only constraining to the diagonal value of $B_m(q, q')$. Beside the diagonal value, the nondiagonal value can have important information, which can be used as the phasing constraint. In the equation 2.20, the only missing information from $B_m(q, q')$ to $I_m(q)$ is the phase for each m. Hence, there is only one unique pieces of information that is unknown.

It is suggested that SVD on the matrix $B_m(q, q')$ will only have one singular value because the matrix $B_m(q, q')$ is a dot product of vector $I_m(q)$ with only the phase missing. The SVD can reveal the number of independent parameter to describe the data. Thus, there will be one singular value of $B_m(q, q')$ because the independent parameter is only the phase of $I_m(q)$. Thus, the previous method in section 4.1 can be improved by constraining

to the nondiagonal value of the $B_m(q, q')$. The expected reconstruction should be much better if the nondiagonal value or SVD is used as a constraint in a polar phasing algorithm.

In section 4.2 the use of triple correlations as additional information to reconstruct electron density is discussed. The derivation of the triple correlation is given from equation 2.21 to equation 2.30. Because of the complexity of the triple correlations, it is used only for sign determination.

The simulation shows that triple correlation and pair correlation can be used to reconstruct electron density from the object that has azimuthal symmetry. By imposing the azimuthal symmetry, only $m = 0$ is nonzero in spherical harmonics expansion. Thus, the magnitude of the $I_{lm}(q)$ can be obtained directly from the diagonal value of $B_l(q, q')$. As a result of that, only the sign of $I_{l0}(q)$ is nonunique and need to be determined from the different information other than $B_l(q, q')$. The nonuniqueness is resolved by trying different signs combination and fitting them to the triple correlations. The set of signs, which is closest to the triple correlations, is taken as the correct combination of the signs. Consequently, the diffraction volume can be constructed from $I_{lm}(q)$ and the electron density is obtainable using a phasing algorithm. This concludes section 4.2 where the triple correlations and pair correlations can be used to reconstruct the electron density from the random angle diffraction patterns.

The explanation and result of how the information about symmetry is obtained from pair correlation are given in section 3.1. Currently, two quantities are used to differentiate the symmetry of the object. Those are the selection rule explained in section 2.3.4 and PCA, which is explained in section 2.4.1. The selection rule is used to differentiate icosahedral symmetry and azimuthal symmetry whereas PCA is used to differentiate azimuthal symmetry, C_n, and asymmetry. The selection rule and PCA complement each other to differentiate the subset of the symmetry. The method suggests that the information of symmetry is not just a mere assumption but also information obtainable from the experiment.

The symmetry determination, which uses the method, requires understanding of the spherical harmonics selection rule and the lowest number of independent parameters of its spherical harmonics expansion. It is possible to extend the determination of the other type symmetry as long as the selection rule and the lowest number of independent parameters are provided. Additionally, currently there is no relation that describe the uniqueness of the symmetry determination. The study of the uniqueness of PCA and the symmetry will complement the theory which I developed.

Another discussion that is described in section 3.1 is the limit of the method. Currently, the inversion symmetry cannot be determined using PCA. The inversion symmetry always exist in reciprocal space. Moreover, the method uses reciprocal space to deduce the symmetry of the object indirectly. As a result of that, the existance of the inversion symmetry of the electron density cannot be determined using method described in section 3.1.

Another application of PCA or SVD of $B_l(q, q')$ is discussed as well. Beside symmetry determination, PCA can be used to check the convergence of $B_l(q, q')$. Section 3.2 explains that some number of diffraction patterns is needed to get the convergence of $B_l(q, q')$. The test that is explained is to check the number of nonzero singular values of $B_l(q, q')$. There is a maximum number of singular values of $B_l(q, q')$ if the $B_l(q, q')$ converges into a form of dot product. The number is $(2l + 1)$, which is the number of singular values for asymmetric structure. Any structure theoretically cannot have more number of singular values more than $(2l + 1)$ because the number of independent parameters to describe asymmetric structure is $(2l + 1)$. In conclusion, if the SVD $B_l(q, q')$ give the number of singular values more than $(2l + 1)$ then the $B_l(q, q')$ doesn't converge.

Section 4.3 explains the reconstruction of the electron density by using pair correlation and positivity constraint. The method uses SVD to get the estimation of $I_{lm}(q)$. The missing or nonunique information is the orthogonal matrix. The orthogonal matrix is determined by imposing the intensity to be positive number. The method defines an

objective function in which if it is zero then the orthogonality is satisfied. The active set algorithm is used to find the zero objective function and at the same time satisfy the positivity constraint. In conclusion, the diffraction volume can be obtained and the electron density is obtained using phasing algorithm.

Currently, the output of the reconstruction is still low resolution reconstruction. The reason for that because there is still a separate step between reconstructing the diffraction volume and the phasing to get electron density. The better reconstruction will be obtained by combining those steps. In other words, it adds additional constraint beside positivity. The constraint comes from any phasing constraint in real space.

Since equation 4.39 relates the intensity to the $O^l_{mm'}$ directly, it is possible to use the relation as additional step in phasing algorithm. It is shown in figure 5.1 how it is done. It involves finding the closest orthogonal matrix or what is known as procrustes problem.

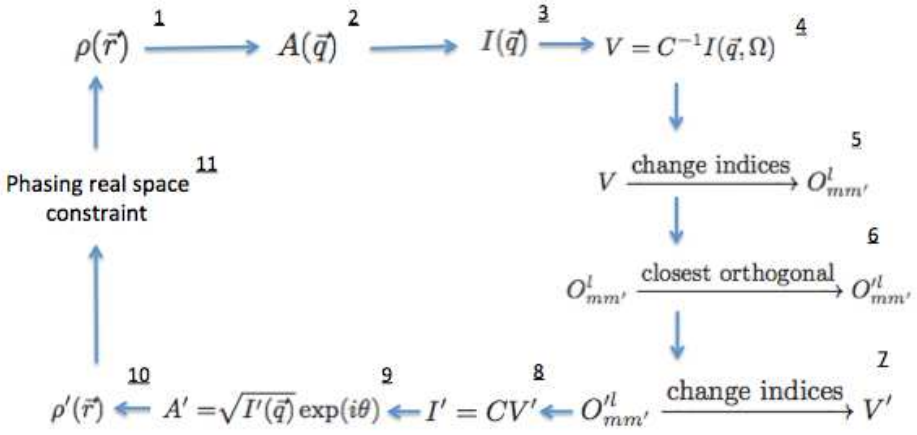


Figure 5.1: Modified phasing algorithm which find closest orthogonal matrix

1. Start initial guess of $\rho(\vec{r})$.

2. Use FFT to calculate $A(\vec{q})$.

3. $I(\vec{q}) = |A(\vec{q})|^2$ and keep information of phase.

4. Estimation of vector V is obtained based on equation 4.39.

5. Change from 1D index of vector V into 3D index of matrix $O^l_{mm'}$ for each l.

6. Find closest orthogonal matrix or it is known as procrustes problem.

7. Change from 3D index of matrix $O^l_{mm'}$ for each l into 1D index of vector V.

8. Use equation 4.39 to obtain next estimation of vector V.

9. Calculate $A(\vec{q})$ from previous information of phase.

10. Use inverse FFT to obtain $\rho(\vec{q})$.

11. Use HIO, ER, shrinkwrap, or charge-flipping to constraint $\rho(\vec{r})$ and cycle is repeated.

The method described above combines the phasing algorithm and the SVD of $B_l(q, q')$ into one iteration. By having it into one iteration, the additional information is obtained from the phasing constraint such as the electron density has to be positive. Thus, it is expected to have a better reconstruction compare to the method that only use positivity constraint.

In conclusion, the main points of this dissertation are:

- The theory for recovering the structure by using an SVD of $B_l(q, q')$ and positivity constraint. The method optimizes an objective function to satisfy $B_l(q, q')$ under the constraint that the diffraction volume is positive. The optimization algorithm is the active set and all quanities are represented in matrix form.

- The demonstration of the possibility of obtaining the information about the symmetry by performing an SVD of $B_l(q, q')$. The symmetry information is obtained by finding the lowest independent parameters for different type of symmetry.

- The theory for reconstructing an azimuthal object from random angle diffraction patterns. The reconstruction uses triple correlation as a constraint to construct diffraction volume.

Appendices

Appendix A

Procrustes Problem

The orthogonal Procrustes problem is defined as finding the orthogonal matrix Ω which transform the matrix A to B or closest to B. Mathematically, it is defined:

$$\Omega A = B \tag{A.1}$$

$$\Omega A - B = 0 \tag{A.2}$$

$$\|\Omega A - B\| = 0$$

$$\text{or} \quad \min\|\Omega A - B\|$$

$$\text{subject to} \quad \Omega^T\Omega = I$$

$$\text{where} \quad \|.\| \quad \text{is Frobenius norm}$$

Frobernius norm can be calculated using trace:

$$\|\Omega A - B\|^2 = \text{trace}(A^TA - 2\Omega^TA^TB + B^TB) \tag{A.3}$$

It is obvious that minimizing the frobernius norm is equivalent to maximizing the trace$(\Omega^T A^T B)$. By decomposing $A^T B$ into its SVD component, then matrix Ω can be determined,

$$\begin{aligned} \text{trace}(\Omega^T A^T B) &= \text{trace}(\Omega^T U \Sigma V^T) \\ &= \text{trace}(V^T \Omega^T U \Sigma) \\ &\leq \sum_i \sigma_i. \end{aligned} \tag{A.4}$$

The trace is maximum if $\Omega = UV^T$ where $[U\Sigma V] = \text{SVD}(A^T B)$.

Appendix B

Active Set Run

The data of run using the active set algorithm is displayed in this appendix. The third column is the objective function and the fourth column is the maximum constraint violation. By the end of iteration, the objective function goes to zero and the violation constraint goes to zero as well. The inputs of the algorithm are the definition of the objective function and the inequality constraint as described in the previous section.

Iter	F-count	f(x)	Max constraint	steplengh	derivative	optimality	Procedure
0	250	251.614	2.43e+08				Infeasible start point
1	500	2728.64	3.659e-09	1	-134	151	
2	750	2554.77	2.179e-09	1	-770	159	
3	1001	351.988	1.153e-09	0.5	-742	117	
4	1252	42.5741	-1.11e-16	0.5	-191	56.1	
5	1504	21.6738	0	0.25	-28.5	5.41	
6	1756	15.2964	0	0.25	-8.09	1.85	
7	2008	11.2793	0	0.25	-5.14	1.86	
8	2260	9.28109	0	0.25	-4.47	1.81	
9	2511	7.78195	0	0.5	-5.09	2.23	
10	2762	7.57845	0	0.5	-6.99	3.39	
11	3014	5.42394	0	0.25	-6.7	1.08	
12	3265	3.94368	0	0.5	-4.21	1.89	
13	3517	2.84324	0	0.25	-4.59	1.01	
14	3769	1.73564	0	0.25	-3.81	0.713	
15	4021	1.19956	0	0.25	-2.38	0.457	
16	4272	1.07737	0	0.5	-1.39	0.531	
17	4523	0.710958	0	0.5	-1.37	0.561	
18	4774	0.533374	0	0.5	-1.26	0.544	
19	5025	0.351412	0	0.5	-1.3	0.406	
20	5276	0.300747	0	0.5	-1.02	0.379	
21	5527	0.183952	0	0.5	-1.11	0.331	
22	5778	0.170421	0	0.5	-0.897	0.304	
23	6029	0.118879	0	0.5	-0.762	0.226	
24	6280	0.0865366	0	0.5	-0.628	0.138	
25	6531	0.0723563	0	0.5	-0.434	0.122	
26	6782	0.0601384	0	0.5	-0.337	0.119	
27	7033	0.0498024	0	0.5	-0.269	0.0627	
28	7284	0.0449978	0	0.5	-0.166	0.0866	
29	7534	0.0440419	9.47e-10	1	-0.145	0.112	
30	7784	0.0438694	4.7e-10	1	-0.224	0.163	
31	8035	0.0306459	5.557e-10	0.5	-0.335	0.095	
32	8286	0.020576	5.409e-10	0.5	-0.303	0.0712	
33	8537	0.0188571	1.291e-09	0.5	-0.177	0.0855	
34	8787	0.0180913	9.951e-10	1	-0.127	0.106	
35	9038	0.0124091	3.236e-10	0.5	-0.185	0.0595	
36	9289	0.00914419	5.582e-10	0.5	-0.17	0.0445	
37	9540	0.00771333	3.364e-10	0.5	-0.107	0.0347	
38	9791	0.00572137	8.294e-10	0.5	-0.0687	0.0314	
39	10041	0.006983	7.911e-10	1	-0.0599	0.0504	
40	10292	0.00438257	5.43e-10	0.5	-0.176	0.0357	
41	10543	0.00328368	6.959e-10	0.5	-0.119	0.0316	
42	10794	0.00304637	1.088e-09	0.5	-0.0641	0.0338	
43	11045	0.00242503	6.596e-10	0.5	-0.067	0.0207	
44	11296	0.00218611	4.362e-10	0.5	-0.0499	0.0178	
45	11547	0.00199401	1.04e-09	0.5	-0.0381	0.0173	
46	11797	0.00221516	9.065e-10	1	-0.0247	0.0252	
47	12047	0.00196579	7.665e-10	1	-0.0574	0.0193	
48	12298	0.00168913	7.255e-10	0.5	-0.06	0.0249	
49	12549	0.00134942	6.616e-10	0.5	-0.0383	0.0103	
50	12799	0.0018409	5.462e-10	1	-0.017	0.0403	

51	13049	0.00118756	7.51e-10	1	-0.0557	0.0214	
52	13299	0.0029403	4.619e-10	1	-0.0435	0.0428	
53	13549	0.00117978	7.498e-10	1	-0.141	0.0191	
54	13800	0.00109385	7.766e-10	0.5	-0.0325	0.027	
55	14051	0.000803772	9.666e-10	0.5	-0.0293	0.0227	
56	14301	0.0013837	8.247e-10	1	-0.0403	0.0369	
57	14551	0.00132486	4.105e-10	1	-0.0721	0.0243	
58	14802	0.00106828	1.546e-09	0.5	-0.0726	0.0322	
59	15052	0.00164059	1.168e-09	1	-0.0381	0.0332	
60	15302	0.00208665	7.705e-10	1	-0.0837	0.0418	
61	15552	0.00269915	4.149e-10	1	-0.0844	0.0574	
62	15803	0.00129852	4.977e-10	0.5	-0.0911	0.0317	
63	16054	0.000794037	8.849e-10	0.5	-0.0534	0.0236	
64	16305	0.000575957	3.842e-10	0.5	-0.0517	0.0163	
65	16556	0.000417857	3.098e-10	0.5	-0.0325	0.00871	
66	16807	0.000394087	7.377e-10	0.5	-0.024	0.00919	
67	17058	0.000334922	4.914e-10	0.5	-0.0128	0.00503	
68	17309	0.000315957	1.319e-09	0.5	-0.0084	0.00541	Hessian modified
69	17559	0.000323258	1.099e-09	1	-0.00577	0.00789	Hessian modified
70	17809	0.000294958	8.416e-10	1	-0.0153	0.00613	Hessian modified
71	18060	0.000275796	1.447e-09	0.5	-0.0121	0.00486	Hessian modified
72	18311	0.000258727	1.017e-09	0.5	-0.0057	0.0032	Hessian modified
73	18561	0.000242878	1.49e-09	1	-0.00349	0.0073	Hessian modified
74	18811	0.000229566	4.976e-10	1	-0.00394	0.00741	Hessian modified
75	19062	0.0001933	1.451e-09	0.5	-0.00493	0.00729	Hessian modified
76	19312	0.000130601	1.869e-09	1	-0.00644	0.00837	Hessian modified
77	19562	0.000281918	1.29e-09	1	-0.00374	0.0209	Hessian modified
78	19812	5.69364e-05	1.632e-09	1	-0.0234	0.00879	Hessian modified
79	20062	0.00038857	1.556e-09	1	-0.00812	0.0142	Hessian modified
80	20312	4.90607e-05	1.259e-09	1	-0.0649	0.00534	
81	20563	0.000151861	3.286e-10	0.5	-0.0115	0.0101	Hessian modified
82	20813	1.38497e-05	2.452e-09	1	-0.0209	0.00535	
83	21063	0.000161841	9.829e-10	1	-0.00751	0.0126	Hessian modified
84	21313	1.57426e-05	1.143e-09	1	-0.0421	0.00351	Hessian modified
85	21564	2.54389e-05	1.061e-09	0.5	-0.005	0.00688	Hessian modified
86	21814	1.8203e-05	1.139e-09	1	-0.0106	0.0048	Hessian modified
87	22065	8.22107e-06	5.164e-10	0.5	-0.00514	0.00422	Hessian modified
88	22316	3.40723e-06	6.661e-10	0.5	-0.00399	0.00179	Hessian modified
89	22567	3.24045e-06	7.577e-10	0.5	-0.00357	0.00245	Hessian modified
90	22817	6.23266e-07	1.312e-09	1	-0.00152	0.00103	Hessian modified
91	23067	4.16197e-06	9.572e-10	1	-0.00185	0.00156	Hessian modified
92	23317	7.60744e-07	9.5e-10	1	-0.00216	0.00107	Hessian modified
93	23567	2.0713e-06	4.351e-10	1	-0.000436	0.00148	Hessian modified
94	23817	1.47535e-06	6.725e-10	1	-0.00149	0.00113	Hessian modified
95	24068	7.18542e-07	3.42e-10	0.5	-0.00123	0.000884	Hessian modified
96	24318	5.57077e-07	2.345e-09	1	-0.000572	0.000766	Hessian modified
97	24568	1.94297e-06	4.621e-11	1	-0.000752	0.00116	Hessian modified
98	24818	6.80905e-07	5.19e-10	1	-0.00125	0.00107	Hessian modified
99	25069	3.89739e-07	1.185e-11	0.5	-0.00181	0.000973	Hessian modified

Appendix C

Protein Data Bank Format

Table C.1: Explanation of the format of pdb file [51]

Protein Data Bank Format: Coordinate Section				
Record Type	Columns	Data	Justification	Data Type
ATOM	1-4	ATOM		character
	7-11	Atom serial number	right	integer
	13-16	Atom name	left*	character
	17	Alternate location indicator		character
	18-20	Residue name	right	character
	22	Chain identifier		character
	23-26	Residue sequence number	right	integer
	27	Code for insertions of residues		character
	31-38	X orthogonal coordinate	right	real (8.3)
	39-46	Y orthogonal coordinate	right	real (8.3)
	47-54	Z orthogonal coordinate	right	real (8.3)
	55-60	Occupancy	right	real (6.2)
	61-66	Temperature factor	right	real (6.2)
	73-76	Segment identifier	left	character
	77-78	Element symbol	right	character
	79-80	Charge		character
HETATM	1-6	HETATM		character
	7-80	same as ATOM records		
TER	1-3	TER		character
	7-11	Serial number	right	integer
	18-20	Residue name	right	character
	22	Chain identifier		character
	23-26	Residue sequence number	right	integer
	27	Code for insertions of residues		character

Appendix D

Cubic Spline

The purpose of this chapter is to derive the parameters in the third order polynomial of the cubic spline function. To simplify the derivation, the x point is represented by parameter t where t is from 0 to 1. The polynomial is represented by,

$$f_i(t) = a_i + b_i t + c_i t^2 + d_i t^3 \qquad i = 0, ..., n-1 \tag{D.1}$$

Based on the boundary condition where the function should be continuous,

$$\begin{aligned} f_i(0) &= y_i = a_i \\ f_i(1) &= y_{i+1} = a_i + b_i + c_i + d_i. \end{aligned} \tag{D.2}$$

Another boundary condition is the first derivative should be continous,

$$\begin{aligned} f_i(0) &= D_i = b_i \\ f_i(1) &= D_{i+1} = b_i + 2c_i + 3d_i. \end{aligned} \tag{D.3}$$

Solving for a_i, b_i, c_i, d_i then gives

$$\begin{aligned} a_i &= y_i \\ b_i &= D_i \\ c_i &= 3(y_{i+1} - y_i) - 2D_i - D_{i+1} \\ d_i &= 2(y_i - y_{i+1}) + D_i + D_{i+1}. \end{aligned} \tag{D.4}$$

The second derivative should also be continuous,

$$\begin{aligned} f_{i-1}(1) &= y_i \\ f'_{i-1}(1) &= f'_i(0) \\ f_i(0) &= y_i \\ f''_{i-1}(1) &= f''_i(0). \end{aligned} \tag{D.5}$$

To have unique solution, another boundary condition is needed. They are second derivative has to be continous,

$$\begin{aligned} f_0(0) &= y_0 \\ f_{n-1}(1) &= y_n \end{aligned} \tag{D.6}$$

A new matrix can be formed based on those constraint. The parameters can be solved

using matrix inversion.

$$\begin{pmatrix} 2 & 1 & & & \\ 1 & 4 & 1 & & \\ & 1 & 4 & 1 & \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & & 1 & 4 & 1 \\ & & & 1 & 2 \end{pmatrix} = \begin{pmatrix} D_0 \\ D_1 \\ D_2 \\ \vdots \\ D_{n-1} \\ D_n \end{pmatrix} \begin{pmatrix} 3(y_1 - y_0) \\ 3(y_2 - y_0) \\ 3(y_3 - y_1) \\ \vdots \\ 3(y_n - y_{n-2}) \\ 3(y_n - y_{n-1}) \end{pmatrix} \tag{D.7}$$

Thus, by inverting the matrix above, the parameters of the interpolation can be determined. In conclusion, the third order polynomial can be used to estimate the value of the function in any point.

References

[1] Neutze, R., Wouts, R., van der Spoel, D., Weckert, D. and Hajdu, J., Nature (London) 406, 752-757 (2000)

[2] LCLS fact sheet 2014, portal.slac.stanford.edu

[3] H. N. Chapman et al., Phil. Trans. Roy. Soc Bi **369**, 20131313 (2014).

[4] N.-T. D. Loh and V. Elser, Phys. Rev. E **80**, 026705 (2009).

[5] R.Fung, V.L. Shneerson, D K. Saldin, and A. Ourmazd, Nature Physics **5** 64i (2009).

[6] A. Barty et al., J. appl. Cryst. **47**, 1118 (2014).

[7] J. C. H. Spence, Private communication.

[8] T. Ekeberg et al. Phys. Rev. Lett. **114**, 089102 (2015).

[9] J. J. Donatelli, P. H. Zwart, and J. A. Sethian, PNAS **112**,i 10286 (2015).

[10] A. Babinet, Compt. Rend, Acac, Sci. **4**, 637 (1837).

[11] Z. Kam, Macromoleculesi **10**, 927 (1978).

[12] Z. Kam, J. theor. Biol. **82**, 15 (1980).

[13] J. Drenth *Priciples of Protein X-Ray Crystallography*

[14] A. Munke, iScientific Data **3**, 16004 (2016).

[15] Xinzheng Zhang, Ye Xiang, David D. Dunigan, Thomas Klose, Paul R. Chipman, James L. Van Etten, and Michael G. Rossmann, PNAS vol. 108 no. 36 (2011)

[16] Borgstahl, G.E., Williams, D.R., Getzoff, E.D., Biochemistry 34: 6278-6287 (1995)

[17] D. K. Saldin, H. C. Poon, P. Schwander, M. Uddin, and M. Schmidt, Optics Express 19, 17318-17335 (2011)

[18] D. K. Saldin, V. L. Shneerson, D. Starodub and J. C. H. Spence Acta Cryst. (2010). A66, 3237

[19] Amand A Lucas and Philippe Lambin, Rep. Prog. Phys. 68 (2005) 11811249

[20] Abramowitz, Milton; Stegun, Irene A., eds. (December 1972) [1964]. "Chapter 9". Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Applied Mathematics Series 55 (10 ed.). New York, USA: United States Department of Commerce, National Bureau of Standards; Dover Publications. p. 355. ISBN 978-0-486-61272-0. LCCN 64-60036. MR 0167642

[21] Cuyt, Annie; Petersen, Vigdis; Verdonk, Brigitte; Waadeland, Haakon; Jones, William B. (2008), Handbook of continued fractions for special functions, Springer, ISBN 978-1-4020-6948-2

[22] Dragia Mitrovi, Darko ubrini (1998). Fundamentals of Applied Functional Analysis: Distributions, Sobolev Spaces. CRC Press. p. 62. ISBN 0-582-24694-6.

[23] A. Jack and S. C. Harrison, On the interpretation of small-angle x-ray solution scattering from spherical viruses J. Mol. Biol. 99, 1525 (1975).

[24] N. V. Cohan "The spherical harmonics with the symmetry of the icosahedral group", Mathematical Proceedings of the Cambridge Philosophical Society 53, 28-38 (1958).

[25] Nandhagopal, N., Simpson, A., Gurnon, J.R., Yan, X., Baker, T.S., Graves, M.V., Van Etten, J.L., Rossmann, M.G. "The Structure and Evolution of the Major Capsid Protein of a Large, Lipid containing, DNA virus" Proc.Natl.Acad.Sci.USA 99: 14758-14763 (2002)

[26] Borgstahl, G.E., Williams, D.R., Getzoff, E.D. "1.4 A structure of photoactive yellow protein, a cytosolic photoreceptor: unusual fold, active site, and chromophore." Biochemistry 34: 6278-6287 (1995)

[27] Stephan Kassemeyer et all, "Femtosecond free-electron laser x-ray diffraction data sets for algorithm development" Optics Express 20: 4149-4158 (2012)

[28] Pendry, J. B., *Low Energy Electron Diffraction* (Academic, London, 1974).

[29] Oszlányi G. and Süto A., Acta Cryst. A **60**, 134-141 (2004).

[30] Oszlányi G. and Süto A., Acta Cryst. A **61**, 147-151 (2005).

[31] Caspar, D. L. D., and Klug, A. Cold Spring Harbor Symp. Quant. Biol. **27**, 1-24 (1962).

[32] Kam, Z., Macromolecules **10**, 927 (1978).

[33] Saldin, D.K., Shneerson, V. L., Fung, R., and Ourmazd, A., J. Phys: Condens. Matter **21**, 134014 (2009).

[34] Tinkham M 2003 Group Theory and Quantum Mechanics (Dover: Courier)

[35] Pande, K., Schwander, P., Schmidt, M., and Saldin, D. K, Phil. Trans. Roy. Soc. B, **369**, 20130332 (2014).

[36] K. Pande, M. Schmidt, P. Schwander, and D. K. Saldin, Structural Dynamics, 2, 024103 (2015).

[37] D. K. Saldin, H. C. Poon, V. L. Shneerson, M. Howells, H. N. Chapman, R. Kirian, K. E. Schmidt, and J. C. H. Spence, Phys. Rev. B 81, 174105 (2010).

[38] P. J. Ho, D. Starodub, D. K. Saldin, V. L. Shneerson, A. Ourmazd, and R. Santra, J. Chem. Phys. 131, 131101 (2009).

[39] H.C. Poon, P. Schwander, M. Uddin, and D. K. Saldin, Phys. Rev. Lett. 110, 265505 (2013).

[40] D. Svergun and H. B. Stuhrmann, New developments in direct shape determination from small-angle scattering 1. Theory and model calculations, Acta Crystallogr., Sect. A: Found. Crystallogr. 47, 736744 (1991).

[41] H. Liu, B. K. Poon, A. J. E. M. Janssen, and P. H. Zwart, Computation of fluctuation scattering profiles via three-dimensional zernike polynomials, Acta Crystallogr., Sect. A: Found. Crystallogr. 68, 561567 (2012).

[42] H. Liu, B. K. Poon, D. K. Saldin, J. C. H. Spence, and P. H. Zwart, Three-dimensional single-particle imaging using angular correlations from x-ray laser data, Acta Crystallogr., Sect. A: Found. Crystallogr. 69, 365373 (2013).

[43] Jeffrey J. Donatelli, Peter H. Zwart, and James A. Sethian, "Iterative phasing for fluctuation X-ray scattering", 1028610291, doi: 10.1073/pnas.1513738112

[44] Jones, T.A., Liljas, L.,Structure of Satellite Tobacco Necrosis Virus After Crystallographic Refinement at 2.5 A Resolution., J.Mol.Biol. 177: 735 (1984)

[45] Charles Kittel, Introduction to Solid State Physics, 2004

[46] D K Saldin, V L Shneerson, M R Howells, S Marchesini, H N Chapman, M Bogan, D Shapiro, R A Kirian, U Weierstall, K E Schmidt and J C H Spence, New J. Phys. 12, 035014 2010

[47] D. K. Saldin, H. C. Poon, V. L. Shneerson, M. Howells, H. N. Chapman, R. A. Kirian, K. E. Schmidt, and J. C. H. Spence, Phys. Rev. B 81, 174105 (2010).

[48] D. T. Cromer, J. B. Mann, X-ray scattering factors computed from numerical Hartree-Fock wave functions, Acta Cryst. (1968). A24, 321-324

[49] INTERNATIONAL TABLES FOR X-RAY CRYSTALLOGRAPHY, MacGillavry, Kluwer Academic Pub

[50] Berman, Helen M. "The protein data bank: a historical perspective." Acta Crystallographica Section A 64.1 (2007): 88-95

[51] Chimera Team, University of California San Diego, "https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/tutorials/pdbintro.html"

[52] Bartels, R. H.; Beatty, J. C.; and Barsky, B. A. "Hermite and Cubic Spline Interpolation." Ch. 3 in An Introduction to Splines for Use in Computer Graphics and Geometric Modelling. San Francisco, CA: Morgan Kaufmann, pp. 9-17, 1998.

[53] Burden, R. L.; Faires, J. D.; and Reynolds, A. C. Numerical Analysis, 6th ed. Boston, MA: Brooks/Cole, pp. 120-121, 1997.

[54] Babinet, A., Compt. Rend. Acad. Sci. 4:638 (1837).

[55] https://github.com/antonbarty/cheetah-old/wiki

[56] Spence J.C.H. private communication.

[57] Drenth, J., Principles of Protein X-Ray Crystallography (Springer Advanced Texts in Chemistry) (Springer: New York, 1994).

[58] Munke, A., et al., Sci. Data, 3: 160064 doi: 10.1038/sdata 2016.64 (2016).

[59] Caspar, D. L. D., and Klug, A. Cold Spring Harbor Symp. Quant. Biol. 27: 1-24 (1962).

[60] Kim, S. S., Wibowo, S., and Saldin, D. K., Internal Medicine Review, in press.

[61] Kurta, R.P., Dronyak, R., Altarelli, M, Weckert, E., and Vartanyants, I. A., New Journal of Physics 15: 013059

[62] Pedrini B., Menzel, A., Guizar-Sicairos, M., Guzenko, V. A., Gorelik, S. David, C., Patterson, B. A., Abela, R., Nature Commun. 14:1647 (2013).

CURRICULUM VITAE

Name: Sandi Wibowo

Place of birth: Jakarta, January 26 1986

Dissertation title: Symmetry and Reconstruction of Particle Structure from Random Angle Diffraction Patterns.

Education :

1. B.Sc University: University of Indonesia (2004-2008)

2. Ph.D University: University of Wisconsin Milwaukee (2010-2016)