The handle http://hdl.handle.net/1887/48877 holds various files of this Leiden University dissertation

**Author:** Li, Y.
**Title:** A new method to reconstruct the structure from crystal images
**Issue Date:** 2017-05-03
A new method to reconstruct the structure from crystal images

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 3 mei 2017
klokke 15:00 uur

door

Yao-Wang Li
geboren te Jiangxi, China
in 1983
Promotiecommissie

Promotores: prof.dr. Jan Pieter Abrahams
co-promotores: dr. Tim Grüne (Paul Scherrer Instituut)

Voorzitter: prof.dr. H.P. Spaink
Secretaris: prof.dr.ir. T.H. Oosterkamp
Overige leden: dr. N.S. Panmu
dr. R.B.G. Ravelli (Maastricht Universiteit)
dr. Xiaodan Li (Paul Scherrer Instituut)
prof.dr. Henning Stahlberg (Universiteit Basel)
prof.dr. T.J. Aartsma

©2017 Yaowang Li
# Contents

## Contents

List of Tables  
List of Figures

## 1 Introduction

### 1.1 Biological macromolecules and cryo-EM

#### 1.1.1 Biological macromolecules

#### 1.1.2 Electron microscopy

#### 1.1.3 Cryo-EM in structural biology

### 1.2 Methods in cryo-EM

#### 1.2.1 Electron crystallography

#### 1.2.2 SPA and ET

### 1.3 Principle of single particle analysis

#### 1.3.1 Image acquisition

#### 1.3.2 Image formation and CTF

#### 1.3.3 Picking up particles

#### 1.3.4 Multivariate statistics analysis

#### 1.3.5 Alignment

#### 1.3.6 Orientation determination and 3D reconstruction

#### 1.3.7 Resolution determination

### 1.4 Principle of electron diffraction

### 1.5 The phase problem in 3D crystals
## Contents

1.6 Outlines of this thesis ........................................ 34

2 Imaging protein three-dimensional nano-crystals ................. 37
  2.1 Abstract ..................................................... 38
  2.2 My contribution ........................................... 38
  2.3 Introduction ............................................... 38
  2.4 Materials and methods ...................................... 41
      2.4.1 Crystallization ........................................ 41
      2.4.2 Vitrification .......................................... 41
      2.4.3 Electron-microscopy data collection .................. 41
  2.5 Results ..................................................... 43
      2.5.1 Structure of the Bragg spots .......................... 43
      2.5.2 Centrosymmetry of the Fourier transform .......... 46
      2.5.3 Visualizing the lattice ............................... 50
  2.6 Discussion and conclusions ................................ 53

3 Filter for processing image data ................................ 59
  3.1 Abstract ..................................................... 60
  3.2 My contribution ........................................... 60
  3.3 Introduction ............................................... 60
  3.4 Method ...................................................... 64
  3.5 Results ..................................................... 67
  3.6 Discussion .................................................. 74

4 A new approach of estimating crystallographic phases ........... 77
  4.1 Abstract ..................................................... 78
  4.2 Introduction ............................................... 78
  4.3 Materials and methods for lysozyme ......................... 81
      4.3.1 Cryo-EM images and diffraction patterns .......... 81
      4.3.2 Forward projection .................................... 81
      4.3.3 Comparison in Fourier space and real space ....... 82
      4.3.4 Reconstruction and validation ....................... 83
  4.4 Materials and methods for peptide VQIVYK ................. 86
      4.4.1 Cryo-EM images and diffraction patterns .......... 86
5 Preliminary results of phasing crystallographic data

5.1 Abstract

5.2 Introduction

5.3 Results and discussion for lysozyme

5.3.1 Reference and experimental projections

5.3.2 Determination of freedom for the experimental projections

5.3.3 Map validation

5.4 Results and discussion for VQIVYK peptide

5.4.1 Contrast improvement using lattice filter

5.4.2 Comparison in Fourier space

5.4.3 Parallel alignment in real space and reconstruction

5.5 Conclusion

6 Discussion, prospects and summary

6.1 Discussion

6.1.1 Data acquisition for 3D protein crystals

6.1.1.1 Detectors in cryo-EM

6.1.1.2 Data collection in cryo-EM

6.1.2 Data processing for 3D protein crystals

6.1.3 The phase problem in 3D electron crystallography

6.2 Prospects

6.2.1 Detectors

6.2.2 Sample holder

6.2.3 Data processing

6.3 Summary

7 Samenvatting
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bash and IMAGIC scripts</td>
<td>131</td>
</tr>
<tr>
<td>References</td>
<td>137</td>
</tr>
<tr>
<td>Curriculum Vitae</td>
<td>157</td>
</tr>
<tr>
<td>Publications</td>
<td>159</td>
</tr>
</tbody>
</table>
List of Tables

1.1 Comparison of electron crystallography and SPA. . . . . 14
## List of Figures

1.1 Growth of gene sequences on the GenBank website (data obtained on 26 September 2016). ........................................ 3
1.2 Growth of released structures per year by experimental method (obtained from the PDB website on 26 September 2016). ........................................................................................................ 4
1.3 Layout of optical components in a basic TEM (Transmission electron microscopy on Wikipedia). .................. 6
1.4 The procedure of electron crystallography: the steps are described in detail. ...................................................... 11
1.5 The procedure of SPA: the steps are described in detail. . 16
1.6 Particles of worm hemoglobin picked from images by the author during the 6th Brazil School for Single Particle Cryo Electron Microscopy. ...................................................... 20
1.7 Averaged images of worm hemoglobin particles calculated by the author during the 6th Brazil School for Single Particle Cryo Electron Microscopy. .............................. 21
1.8 Flowchart of angular reconstitution and 3D reconstruction. 25
1.9 Interaction of electrons with atoms in matter. ............... 29
1.10 The Laue equation. .......................................................... 30
1.11 Bragg’s law. ................................................................. 31
1.12 The Ewald sphere. ......................................................... 32
2.1 Lysozyme nano-crystals in a crystallization drop imaged with a light microscope. ................................................. 42
List of Figures

2.2 The shape of the crystal can be showed after calculating the local variance and its lattice contrast can be enhanced by a Wiener type filter. .............................................. 44
2.3 Fourier transformation shows the lattice spots of crystal. . 45
2.4 Histogram of the maximum resolution observed in electron images of 200 different lysozyme three-dimensional nanocrystals. ................................................................. 46
2.5 Two defocused frames from raw data of a tomographic series of one of the crystals shows its thickness. ................. 50
2.6 Electron diffraction pattern showing lunes (at 200 keV). . 52
2.7 The patches in the same domain can be averaged to enhance the contrast. ...................................................... 54
2.8 Examples of averaged images with high contrast. ............. 55

3.1 The measured structure factor $F_m(h)$ is the sum of the unknown structure factor corresponding to the lattice signal $F_l(h)$ and the unknown structure factor corresponding to the noise $F_n(h)$. .................................................. 64
3.2 Plot of the rotational average of fig. 3.3. ......................... 65
3.3 Images of 100 nm thick 3D crystals, produced by a Titan Krios 300 kV FEG transmission EM, captured with a Falcon 2 camera (4096×4096 pixels) using an 0.5 sec exposure time and an illumination of 3 to 10 $e^-/A^2$. ................. 68
3.4 Rotational average which is subtracted from the power spectrum during lattice filtering. ................................. 70
3.5 Unique half of the centro-symmetric lattice filter of the top lysozyme crystal image. ................................. 71
3.6 Unique half of the centro-symmetric lattice filter of the top lysozyme crystal image. ................................. 72
3.7 Results from the same area as in figure fig:original-proccessed and results. .................................................. 73

4.1 The flowchart of reconstructing map-3. ......................... 83
4.2 The flowchart of creating map-2. ......................... 84
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3</td>
<td>The flowchart of creating $map-1$.</td>
<td>85</td>
</tr>
<tr>
<td>5.1</td>
<td>The CC value curves in Fourier space.</td>
<td>95</td>
</tr>
<tr>
<td>5.2</td>
<td>The experimental projections and reference projections were compared in Fourier space and real space.</td>
<td>96</td>
</tr>
<tr>
<td>5.3</td>
<td>The FSC curves between the four maps (three reconstructed maps and the original map).</td>
<td>100</td>
</tr>
<tr>
<td>5.4</td>
<td>The FSC curves of $map-4$ with $map-0$ and $map-1$.</td>
<td>101</td>
</tr>
<tr>
<td>5.5</td>
<td>The two $\beta$ distributions in the Euler angles.</td>
<td>101</td>
</tr>
<tr>
<td>5.6</td>
<td>The orthogonal views of the reconstructed maps and the original map.</td>
<td>102</td>
</tr>
<tr>
<td>5.7</td>
<td>Noise was suppressed by the lattice filter.</td>
<td>103</td>
</tr>
<tr>
<td>5.8</td>
<td>The experimental projections and reference projections were compared in Fourier space.</td>
<td>105</td>
</tr>
<tr>
<td>5.9</td>
<td>The CC value curves in Fourier space.</td>
<td>106</td>
</tr>
<tr>
<td>5.10</td>
<td>The FSC curve of the map reconstructed from the experimental projections for peptide VQIVYK.</td>
<td>107</td>
</tr>
<tr>
<td>5.11</td>
<td>The orthogonal views of the reconstructed map of the peptide.</td>
<td>108</td>
</tr>
</tbody>
</table>
List of Figures
Chapter 1

Introduction
1. Introduction

1.1 Biological macromolecules and cryo-EM

1.1.1 Biological macromolecules

Living cells contain four types of macromolecules, which play an important role in the processes of life. They are carbohydrates, lipids, proteins, and nucleic acids. Carbohydrates can be used as fuel and building material, and they are the main source of immediate energy. Lipids mainly store a large amount of energy, and they can also protect the cellular contents. Proteins have many types of structures, which have different functions. These functions can be structured support, storage, transport, cellular signaling, movement, and defense against foreign substances. Nucleic acids are of two types: ribonucleic acid (RNA) and deoxyribonucleic acid (DNA). They are important because they are specialized for storage and transmission of inherited information between generations [1].

Proteins are the building blocks in a cell, and a typical cell may contain thousands of different proteins with different functions. For example, membrane proteins can communicate and interact with the environment or with different compartments. Solving their structures is the first vital step towards understanding their functions. The structural information of proteins can be used for drug design or other purposes. Membrane proteins represent the majority of drug targets in pharmaceutical research, and 50% of all modern therapeutics target G protein-coupled receptors [2]. However, the rate of obtaining genomic sequences has been much faster than that of determining structures. According to the data in GenBank and PDB on 26 September 2016, over 196 million genomic sequences have been obtained, while only 122799 macromolecules (113999 proteins) have been determined (see Fig. 1.1). X-ray, NMR and electron microscopy are the experimental methods for determining the structures. X-ray especially is the main method, and has solved most
of the macromolecules. Since 2013, however, electron microscopy has developed dramatically, with the number of macromolecules determined by this method increasing, while the number determined by X-ray is decreasing, although X-ray is still the routine mainstream method for structure determination (see Fig. 1.2).

![Figure 1.1: Growth of gene sequences on the GenBank website (data obtained on 26 September 2016).](image)

So why has X-ray crystallography lost its advantage over electron microscopy? The answer is simple: radiation damage. Many biological or organic molecules cannot survive in a strong radiation environment, because they cannot grow into crystals of sufficient size, or can hardly grow at all. Electron microscopy, including electron crystallography and single particle analysis (SPA), has considerable advantages. Electrons have a stronger interaction with atoms (more elastic events) and deposit less energy in the sample (less radiation damage) than X-rays. Consequently, much smaller crystals can be analyzed with electron crystallography. In addition, electrons can be focused by magnetic lenses to form an image, which means the phase information lost in X-ray crystallography is preserved in images. It is therefore feasible to investigate the structure of biological molecules using electron microscopy: either electron crystallography or SPA.
Figure 1.2: Growth of released structures per year by experimental method (obtained from the PDB website on 26 September 2016).
1.1.2 Electron microscopy

Microscopy has evolved from light to electrons, since it became necessary to have atomic detail. Electrons, like X-rays, have a short wavelength, and can achieve resolution of $10^{-10}$ m. This strong resolving power is sufficient to show atomic detail. The wavelength of electron radiation in a transmission electron microscope (TEM) is at the subangstrom level, for example, the wavelength of electrons at 100 kV is about 0.04 Å [3].

Louis de Broglie (1925) was the first to propose that electrons have wave-like characteristics, and that their wavelength is less than that of visible light. In 1927, electron diffraction experiments conducted independently by two research groups, Davisson and Germer and Thomson and Reid, demonstrated the wave nature of electrons. An important breakthrough was then made in 1932: the electron lens was developed into practical reality, and the first electron microscope was reported by Knoll and Ruska. Following on from this, the Cambridge group developed the theory of electron diffraction contrast [4].

1.1.3 Cryo-EM in structural biology

Initially, the resolving power of electron microscopy did not work on biological molecules, because radiation damage was caused and the images also had a low signal-to-noise ratio (SNR), until a new sample preparation technique appeared. This was the flash-freezing technique, where the sample is frozen at the temperature of liquid nitrogen. This technique was reported by Jacques Dubochets group at EMBL, who had discovered how to rapidly freeze very thin films of molecules, so that the molecules were embedded in amorphous ice at this temperature [5]. The cryo-EM technique provides a way to observe the real native state structure as it exists in solution, by freezing the samples extremely fast in a layer of vitreous ice [6]. Freezing reduces electron radiation damage by keeping it localized and therefore limited. It also prevents
1. Introduction

![Diagram of TEM Components](image)

**Figure 1.3:** Layout of optical components in a basic TEM ([Transmission electron microscopy](https://en.wikipedia.org/wiki/Transmission_electron_microscopy) on Wikipedia).

---

6
the evaporation of water from the molecules as the sample is imaged in the TEMs vacuum. Cryo-EM is therefore the obvious choice for studying large biological complexes, and its tremendous potential in biological structure determination was already recognized in the early 1990s [7].

The flash-freezing technique makes it possible to reduce the radiation damage and to increase the samples survival time, although it is still difficult to collect enough data in a limited time using the traditional methods. Fortunately, the cryo-EM technique has improved, and a growing number of structures have been solved by this method. The greater capability of cryo-EM in recent years is mainly due to the improved microscopes, especially the new direct electron detectors (DED), which are based on complementary metal oxide semiconductor (CMOS) technology [8; 9], and image processing algorithms. The detective quantum efficiency (DQE) of the new detectors is much higher at 300 keV than the previously best traditional film [10], and the increased DQE can obviously improve the SNR of the images. The principle of the DED is different from that of the charge-coupled device (CCD): whereas the CCD first converts the electrons to photons of light, the DED directly counts the electrons. It is therefore able to read out data faster and to have a movie mode, which offers a new way to improve SNR by correcting the blurred images that result from tiny beam-induced movements of the sample [9].

There are mainly three methods in electron microscopy: SPA, electron tomography (ET), and electron crystallography. Using SPA, Liao et al. recently solved the transient receptor potential ion channels at 3.4 Å [11]. Another example is the 6 Å resolution of the HIV-1 envelope glycoprotein structure by Mao [12]. The resolution of these solved structures is at the near-atomic level. With electron crystallography, it is relatively easy to achieve atomic resolution. Since the first structure of bacteriorhodopsin was solved using electron crystallography [13], much progress has been made with improving this method and achieving high resolution. Many structures have been solved at atomic resolution level,
1. Introduction

for example mammalian AQP0 at 1.9 Å [14].

All in all, the sample preparation, low-dose radiation, new microscopes and new detectors work together to improve the SNR of images resulting from radiation damage. The algorithms for image processing and automation capabilities also contribute to the molecular resolution [15]. It is therefore reasonable to suppose that cryo-EM will become the mainstream technology for research on biological molecules.

1.2 Methods in cryo-EM

SPA, ET and electron crystallography are the main methods in cryo-EM. SPA and ET are used for molecules in solution, while electron crystallography is used for crystals. Electron crystallography mainly solves membrane proteins, which are easier to grow into two-dimensional (2D) crystals than three-dimensional (3D) crystals. Membrane proteins in 2D crystals are in a native environment, which is a lipid bilayer. All 3D protein crystals, not only the membrane proteins, are usually very small, such as micro-crystals or nano-crystals. They cannot be analyzed by X-ray crystallography. Developments have been made in relative techniques, especially microscopy and detectors, for these 3D crystals and some progress has been achieved. SPA usually analyzes biological molecules that cannot be crystallized and are isolated in solution. SPA therefore has a great advantage and can extend the research range of biological samples.

1.2.1 Electron crystallography

Electron crystallography was established on the basis of the research work on bacteriorhodopsin in 1975, in which the structure was solved from images and diffraction patterns of 2D crystals [16]. Electron crystallography was first applied to 2D crystals, mainly those of membrane proteins. This was because crystalline arrays of proteins are typically
just a single molecule thick, which means that high resolution structures can be achieved through image processing with Fourier filtering and averaging to enhance the SNR and contrast. The better the order of the 2D crystal, the higher the resolution that can be achieved. In addition, electrons can be focused by magnetic lenses, which makes it possible to image crystals and obtain phase information directly. This technique has been used to successfully determine the structures of several proteins from thin 2D crystals [17] and can solve the protein structures at atomic resolution [14]. Nevertheless, high energy electrons result in a large amount of radiation damage to the sample, leading to a loss of resolution and destruction of the crystalline material [18]; this cannot be avoided. In order to solve a structure, many 2D crystals have to be tilted, because each crystal can usually yield only a single diffraction pattern. Hence, all the data from many individual crystals need to be merged.

Data processing in 2D electron crystallography mainly involves the processing of images and diffraction patterns 1.4. This must be done before they can be merged to obtain high resolution data, as the amplitude of the Fourier order of the images is incorrect and there is no phase information in the diffraction patterns. For the images, the contrast transfer function (CTF) of the electron microscope makes large changes to the amplitude where information was lost in the region of CTF zero crossings, while the phase shifts can be accurately corrected. Highly tilted 2D crystals will lead to some areas being imaged at extreme defocus values and therefore suffering from resolution loss. However, electron diffraction patterns are insensitive to sample movements and variation in defocus, therefore amplitude information from the diffraction patterns has high resolution.

The processing of images requires pre-processing to be performed, with the aim of sample movement correction and CTF correction. The DED makes it possible to obtain data quickly, which means that the movements of the sample can be recorded. Cross correlation (CC) can
1. Introduction

determine the movements and correct them. Meanwhile, the CTF also needs to be corrected, because these images were modulated by the microscope. The next step is 2D processing of the images, which aims to extract the amplitude and phase of the Fourier order. This involves the tilt geometry, lattice indexing, crystal symmetry, and crystal unbending. The tilt geometry must be obtained by measuring the local defocus via a Thon ring, in which the tilt angle, tilt axis, and angle between the principle lattice vector and the tilt axis have to be measured. These will help to calculate its vertical height in reciprocal space. We then need to index the lattice, which is defined by two basic vectors. Next, the crystal symmetry has to be determined; only 17 symmetry groups exist in 2D crystals. The following step is to correct the distortion. Images of 2D crystals suffer from out-of-plane bending, caused by bent carbon films, and the unbending of these images attempts to position each crystal unit cell in its ideal lattice location. This distortion vector information can be obtained from the CC map between a reference and the Fourier filtered version of the image [19; 20].

For the processing of diffraction patterns, the beamstop position and shape are determined first and then a mask is made of the beamstop. Next, the background caused by the inelastically scattered electrons and the elastically scattered electrons from noncrystalline parts of the sample needs to be subtracted. After this, lattice determination for the diffraction patterns must be done; this is more difficult than lattice determination of the images, because the origin of the lattice is not in the center of the image, and both the lattice origin and low resolution diffraction spots are covered by the beamstop. Finally, the tilt geometry also has to be calculated, using only the lattice distortion, as there is no defocus gradient [19; 20].

These data, both images and diffraction patterns, need to be merged into 3D reciprocal space. They are combined according to the central section theorem, which states that they are a central
section through the Fourier transform of a 3D unit cell. The different images are aligned to find the common phase origin. The amplitudes and phases of the crystals are concentrated on the lattice in the $z^*$ direction. These data are then scaled and the inverse Fourier transform is performed to obtain the density map of the unit cell [19; 20].

Figure 1.4: The procedure of electron crystallography: the steps are described in detail.

For 3D electron crystallography, some factors that must be considered are the beam sensitivity, the large unit cell and dynamical scattering. This last factor is caused by the thickness of the crystal, which has more than one unit cell in the third axis. This therefore increases multiple scattering and nonlinear effects in electron diffraction and imaging. This dynamical scattering problem of small molecules has been tackled with multi-slice least-squares methods [21]. Progress was
made first on inorganic 3D micro- and nano-crystals. Two methods that can automatically collect 3D diffraction data from a single 3D micro- or nano-crystal have been developed: automated diffraction tomography (ADT), developed by Kolbs group [22; 23; 24], and rotation electron diffraction (RED), developed by Hovmiøller and Zous group [25; 26; 27]. Both methods can be used for collecting almost complete 3D electron diffraction data. ADT normally uses discrete goniometer tilt, while RED combines goniometer tilt and electron beam tilt. RED has the advantage that data collection can be controlled entirely by software and performed on a conventional transmission electron microscope without any additional hardware. The second area of progress was in controlling radiation damage of organic crystals. Preserving the sample in amorphous ice and low-dose electron radiation work together to delay radiation damage. Meanwhile, new detectors were developed, which make it possible to collect data faster. In 2013, our group succeeded in collecting rotation diffraction data from a single 3D nano-crystal using the Medipix2 detector, which is a quantum area detector [28]. In the same year, Dan et al. solved the lysozyme structure from a single 3D micro-crystal. In their research, in addition to using a sensitive CMOS based detector, an important factor is extremely low-dose radiation; each exposure lasted up to 10 seconds at a dosage of approximately 0.01 $e^-/\AA^2$, with the cumulative dose less than 9 $e^-/\AA^2$. They named this method microED [29]. The next year, the structure of bovine liver catalase was determined from a single crystal at 3.2 $\AA$ resolution using this method [30]. In 2015, Yonekura solved the structures of $Ca^{2+}$-ATPase and catalase at resolutions 3.4 $\AA$ and 3.2 $\AA$ respectively with microED [31]. In 2014, microED was improved with a new data collection protocol that can yield more accurate data; in the original protocol, data were taken as a tilt series of still exposures, while in the new protocol electron diffraction data are collected in the form of a movie, as the crystal is continuously rotated by the microscope. These data can then be processed directly by the well-established program for X-ray crystallography (MOSFLM) [32].
1.2.2 SPA and ET

SPA and ET are two methods that try to reconstruct the structure of a molecule from 2D projections [33]. They have benefited from improved microscopes, detectors, and software, which can dramatically improve the SNR of images. This makes it feasible to reconstruct near-atomic resolution structures from noncrystalline samples. SPA mainly works on the 2D projections of macromolecules, while ET is mostly used for sub-cellular structures. The underlying principle is that numerous molecules which are imaged in different orientations can be aligned, averaged and reconstructed into a 3D density map. The recorded images of individual molecules (single particles) need to be processed so that their CTF parameters are estimated and corrected in order to recover information. Then, particles are picked up and analyzed using multivariate statistics analysis (MSA), and averaged to obtain better SNR class images. The following step is to determine their orientations for reconstruction.

Determining the orientations is one of the most important steps. The method for determining the orientations of 2D projections is different for SPA and ET. The classical tomography approach requires many pictures of the same particles tilted into different orientations, which is difficult because of the radiation damage mentioned previously [34; 35]. The technique has benefited from the introduction of automated tomography [36]. There are also two main problems relating to resolution. One is that the maximum tilt of the specimen stage is restricted by the sample holder and other factors, resulting in a loss of resolution in the vertical direction. The second is that resolution in the plane is also poor, because of the method of data collection. This problem has been resolved by collecting data using double-tilt tomograms. Another tilting approach requiring only two exposures of the sample is the random conical tilt
1. Introduction

Electron crystallography SPA

<table>
<thead>
<tr>
<th>Advantages</th>
<th>SPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>well-established techniques and software</td>
<td>not necessary to crystallize</td>
</tr>
<tr>
<td>can handle nano-size crystals</td>
<td>there is no phase problem</td>
</tr>
<tr>
<td>strong diffraction with matter at atomic resolution</td>
<td>needs less materials</td>
</tr>
<tr>
<td>highest atomic resolution structure achieved</td>
<td>easy for large molecules</td>
</tr>
<tr>
<td>molecule is in the native state</td>
<td>too much computational cost</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Disadvantages</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>difficult to grow crystals</td>
<td>lower resolution</td>
</tr>
<tr>
<td>difficult to solve in presence of disorder</td>
<td>less developed for different conformational states</td>
</tr>
</tbody>
</table>

Table 1.1: Comparison of electron crystallography and SPA.

(RCT) [37]. Although it is not difficult to obtain orientations of 2D projections for either ET or RCT, multiple exposures are required. SPA, which uses a single exposure, can intrinsically reach high resolution. Well-developed methods for orientation determination are projection matching and angular reconstitution [38; 39]. Projection matching is based on searching for the projection orientation of individual molecular images by calculating the correlation of images and re-projections of an earlier 3D reconstruction [40]. Harauz and Ottensmeyer succeeded in orienting individual nucleosomes for 3D reconstructions relative to an earlier model structure [41]. In addition to nucleosomes, orientations of icosahedral and asymmetric structures have also been determined by the projection matching technique [42; 43]. Compared with projection matching, angular reconstitution is more general and powerful, and this technique is based on searching for the common line projections in 2D projections of a 3D structure. Angular reconstitution has already yielded structures at resolution levels of Å [40]. With the new direct electron detectors, this resolution could be enhanced to 3 Å [44].

Collectively, electron crystallography and SPA have their own advantages and disadvantages as listed in Table 1.1.
1.3 Principle of single particle analysis

SPA reconstructs the 3D structure of a macromolecule from a set of cryo-EM projections. The basic concept of single particle reconstruction dates back to the early 1970s, when the first 3D reconstructions of the human wart virus and the bushy stunt virus were acquired using images of fewer than 10 particles [45; 46]. Macromolecules exist as many isolated particles in the micrograph, ideally distributed randomly in a layer of vitreous ice. This method needs thousands of images of molecules to rebuild its 3D structure. The procedure of SPA is shown in Fig. 1.5.

1.3.1 Image acquisition

Specimen preparation is the first and most important step. There are a number of basic requirements for specimens that are used for high resolution single particle reconstruction. Here, art meets science. The method of specimen preparation must stabilize the hydrated molecule so that it can be observed in the vacuum environment of the EM. Another issue is the contrast problem: the contrast produced by the molecule itself is weak. Negative staining with heavy metal salts can improve the contrast, but then only the surface of the molecule can be imaged and the internal information is lost in the reconstructed map. If images are obtained under focus, this will also enhance the contrast. Because the images have been modulated by the CTF, which depends on defocus settings, and the amplitudes of the high spatial frequencies are reduced by the envelope function of the CTF, a higher defocus enhances the low resolution image contrast while decreasing the high resolution contrast. This means that a high defocus limits the frequency range of useful information. It is therefore better to use the lowest possible defocus that can generate sufficient low resolution contrast for picking up particles. The amplitude of the image can be increased by also using the
1. Introduction

Figure 1.5: The procedure of SPA: the steps are described in detail.
objective aperture, which can block electrons scattered at high angles. However, it also will cut off the high resolution information, therefore the largest possible objective aperture is needed. In addition, increasing the electron dose can boost the contrast, although biological molecules are beam sensitive, and can easily be destroyed by radiation damage [47].

In fact, radiation damage is the fundamental problem. In order to decrease radiation damage, scientists have tried to embed the molecules in different mediums that closely approximate to the aqueous environment, and to collect images produced by the molecule itself. The medium could be glucose [16], tannic acid [48], vitreous ice [49; 50; 51; 52] or cryo-negative staining [53]. Vitreous ice can delay the radiation damage, and it is clear that specimens embedded in vitreous ice are the best for preservation of interior features [54].

1.3.2 Image formation and CTF

The image collected from microscopy is actually a distorted version of the projection density of the molecules, so it is important to know the imaging conditions and the principle of the underlying theory, namely as CTF. This theory can explain how the image was formed, although it ignores a number of effects whose magnitudes vary from one specimen to another (e.g. inelastic and dynamical scattering, Ewald sphere curvature) [54]. In other words, CTF is an approximation to the comprehensive theory of image formation [55].

When an electron hits an atom, both elastic and inelastic scattering can occur. Elastic scattering contains high resolution information, while inelastic scattering produces noise. The elastic scatter can be described as a phase shift $\Phi(r)$. This means that the wave function $\Psi_{bf}$ in the back focal plane is the Fourier transformation of the wave plane exiting the sample. In addition, the lens aberrations and the defocus can also shift the phase ($\chi(k)$) [54].
1. Introduction

\[ \Psi(r) = \Psi_0 e^{i\Phi(r)} \]  \hspace{1cm} (1.1)

\[ \gamma(k) = 2\pi \chi(k) \]  \hspace{1cm} (1.2)

\[ \Psi_{bf}(k) = F(\Psi(k)) e^{i\gamma(k)} \]  \hspace{1cm} (1.3)

where \( r \) is a column vector \([xy]^T\). We suppose that the incoming wave is traveling in the z-direction, \( \Psi_0 \) is the incoming wave plane and \( \Psi(r) \) is the wave plane after its phase was shifted by a term \( \Phi(r) \), caused by the elastic scatter occurring in the sample; \( k \) defines the coordinates in the back focal plane and \( \chi(k) \) is the wave aberration function due to the lens aberrations; \( \Psi_{bf} \) is the Fourier transform (F) of the wave function behind the object, multiplied by a term that represents the effect of phase shift [54].

In other words, the wave function in the image plane is the inverse Fourier transform (\( F^{-1} \)) of the wave function \( \Psi_{bf} \) after this wave function has been modulated by an aperture function \( A(k) \) that blocks part of the wave function [54].

\[ \Psi_i(r) = F^{-1}(\Psi_{bf} A(k) e^{i\gamma(k)}) \]  \hspace{1cm} (1.4)

\( A(k) \) is actually a rectangular function and gives a pulse to the wave function; it equals 1 if \( |k| = \theta/\lambda \leq \alpha/\lambda \), otherwise it is 0. Here \( \theta \) is the scattered angle and \( \alpha \) is the angle corresponding to the radius of the objective aperture [54].

Finally, there is the observed intensity distribution \( I(r) \) in the image plane. If we ignore irrelevant scaling factors, it is the square of the absolute value of \( \Psi_i(r) \) [54].
\[ I(r) = |\Psi_i(r)|^2 \quad (1.5) \]

We now know the image formation in electron microscopy, and the combined effect of lens and aperture that modulates the wave function: the phase CTF. The defocus value also needs to be determined for CTF correction in order to improve the resolution. The CTF correction can be done either before or after particles have been picked (identified) from the electrographs [54]. There are a number of methods that can be used to correct the CTF [56; 57; 58; 59; 60; 61].

\[ A(k) e^{i\gamma(k)} \quad (1.6) \]

### 1.3.3 Picking up particles

Particles first need to be extracted from images, as they will be classified into different groups based on their orientations. There are two different methods to accomplish this task: by hand (particles can be manually selected on the image) and by an automatic procedure. Most programs, such as IMAGIC [62; 63], Spider [64; 65], and EMAN2 [66; 67], for instance, allow both methods for picking up particles. Nowadays the automatic procedure is especially used when we have a whole stack of micrographs. Particle selection can be carried out in several ways in IMAGIC, using the reference free option variance/modulation or the cross-correlation function (CCF)/mutual-cross function (MCF) correlation to search for particles that matched a given (set of) reference(s). Here the variance/modulation means that the input images are converted into variance/modulation images as discussed by Van Heel in 1982 [68], in which particles show up as peaks of high local variance or modulation. For the correction methods, we need a (set of) reference(s), which is normally generated from an existing model or a previous 3D reconstruction if we want to select particles automatically. However, references can also
1. Introduction

be generated by picking up a set of particles by hand. The CCF/MCF means that we use CCF or MCF between a set of input images and reference images. Frank and Wagenknecht developed a method based on CC to search for a radially averaged reference image showing the molecule in a selected orientation [69]. The result of cross correlating the micrograph with such a disk with the approximate size of the particle is analyzed. Peak search, then, yields a list of center coordinates of particle candidates stored in a document file. Fig. 1.6 shows some particles of worm hemoglobin. These particles obviously display a very low contrast.

Figure 1.6: Particles of worm hemoglobin picked from images by the author during the 6th Brazil School for Single Particle Cryo Electron Microscopy.
1.3.4 Multivariate statistics analysis

The particles have been picked up from images with a low contrast, and their orientations are random (Fig. 1.6). We can improve the SNR and contrast by means of MSA [68; 70; 71]. This can divide the particles into different subgroups and then average them to calculate the average image or class image for each subgroup, which dramatically boosts the SNR (Fig. 1.7).

![Figure 1.7: Averaged images of worm hemoglobin particles calculated by the author during the 6th Brazil School for Single Particle Cryo Electron Microscopy.](image)

The averaged particles are then classified into subgroups using principal component analysis (PCA). In order to analyze a set of images with PCA, these images need to be represented in a different way. We suppose that there is a hyperspace \( \mathbf{X} \), in which the number of dimensions is the number of pixels in the image. Then, each image is
regarded as a point in this hyperspace. The distribution of these points gives important information: how many domains can be obtained and how many images there are in each domain. Images in the same domain can be averaged to obtain a high contrast averaged image. These averaged images can be used as references to align the original images again [54].

\[
X = \begin{pmatrix}
x_{11} & x_{12} & \ldots & x_{1J} \\
x_{21} & x_{22} & \ldots & x_{2J} \\
\vdots & \vdots & \ddots & \vdots \\
x_{N1} & x_{N2} & \ldots & x_{NJ}
\end{pmatrix} \tag{1.7}
\]

Here, \( N \) is the number of images or vectors in hyperspace, and \( J \) is the number of dimensions [54].

Before we calculate the eigenvectors, we need to do some pre-processings of this matrix (\( X \)): first we normalize each row, and then we calculate the self-correlation.

\[
D = (X - \bar{X})' (X - \bar{X}) \tag{1.8}
\]

\[
Du = \lambda u \tag{1.9}
\]

Here, \( \bar{X} \) is a matrix containing the average image in each row, \( D \) is the covariance matrix, \( \lambda \) is a multiplier, and \( u \) is the eigenvector [54].

Another method of classification is correspondence analysis (CA), which is similar to PCA. The difference between the two methods is the way of calculating the distance: CA is based on the computation of \( \chi^2 \) distance, while PCA is based on the Euclidean distance [54].
1.3.5 Alignment

There is an important technique that can improve the SNR of class images: the alignment technique. We can align images to a reference image or directly to each other, using a correlation function (CF). The conventional correlation function is the CCF [72] as described above, which reveals the translation vector and rotation matrix relating two different images. If we calculate the CCF between an image and itself, then we call it the auto-correlation function (ACF). Although the CCF is defined in real space, calculating the CCF in real space is very expensive in terms of CPU usage. The CCF is therefore calculated in Fourier space instead of real space, using the efficient fast Fourier transform (FFT) algorithm. Supposing we have one particle (p) image and one reference (r) image, then

$$CCF = F^{-1}(F(p) \ast F(r)) \quad (1.10)$$

Here, $F$ represents the Fourier transform and $F^{-1}$ is the inverse Fourier transform.

This procedure is equivalent to the convolution theorem: the convolution of two functions is equal to the inverse Fourier transform (FT) of the product of their Fourier transforms. To calculate the correlation, rather than the convolution, the CCF first calculates the FT of each of the images concerned. Then, the transforms are conjugate-multiplied together and the result is transformed back to real space, resulting in the CCF. The peak value in CCF corresponds to the translation that maximized the correlation between the images [54].

Another CF is the MCF that was introduced by Van Heel in 1992 [73]. It is similar to CCF, but with an amplitude-square-root (ASR) filtered version of the image in CF. In practice, both images are first transformed into Fourier space, after which we take the square root of the amplitudes of each Fourier space pixel before the complex conjugate
1. Introduction

multiplication of the two transforms. Finally, the inverse Fourier transform yields the MCF. MCF can avoid a problem that occurs in CCF, namely that strong frequency components in the images have an overwhelming influence on the results [74].

Hence, the alignment of the data set of particles to a set of references can use either CCF or MCF. It is relatively easy to determine the shift vector, calculate the CCF between two images and search for the highest value in the CCF map. However, it is more complex to calculate the rotational angle. In fact, determination of the rotational angle is similar to determination of the shift vector. The difference is that one-dimensional (1D) angular projections (through radon transform) of images, rather than the images themselves, are used for calculating the CCF map, and the maximum peak value is the corresponding rotational angle.

1.3.6 Orientation determination and 3D reconstruction

Because these class images are randomly oriented in 3D space, determining their orientations is one of the most important steps for 3D reconstruction. If an initial 3D structure exists, then projection matching can be used to calculate or improve the orientation of the class averages.

Another method for determining orientations is angular reconstruction, which is based on the common line theorem. This theorem is similar to the one on which 3D reconstruction is based: the central section theorem [75]. The common line and central section concepts actually belong to the same theorem but in different dimensions. Fig. 1.8 shows the relationship between them and also their relationships in real space and Fourier space [76]. The central section theorem states that
the Fourier transform of any 2D projection of a 3D density distribution is a 2D section through the center of the 3D Fourier transform of the 3D density distribution. The common line theorem states that a 1D projection of a 2D density distribution is equivalent to a 1D central line through the 2D Fourier transform of the 2D density distribution; two central sections have one central line in common [76].

\[ \text{Figure 1.8: Flowchart of angular reconstitution and 3D reconstruction [76].} \]

In other words, the angular reconstitution approach is based on the fact that two different 2D projections of a 3D object always have a 1D line projection in common, and from the angles between such common line projections, the relative Euler angle orientations of projections can be determined. For asymmetric objects, a minimum of three projections are required, which should not be related by a tilt around a single rotation axis [39].
1. Introduction

Once the orientations have been determined, a 3D structure model can be reconstructed. In the IMAGIC software, the main 3D reconstruction algorithm is currently the exact filter back-projection algorithm. In filter back-projection techniques, the Fourier space filtering is generally performed using an analytical filter. This filter is implicitly based on the a priori assumption that an infinite number of 2D projections of the 3D structure are available and that their projection directions are uniformly distributed over all possible angles [63].

1.3.7 Resolution determination

Resolution measures in cryo-EM evaluate the consistency of results in reciprocal space, which indicates the overall quality of the experiment and the uniformity of the reconstructed structure. An objective method does not exist. The Fourier shell correlation (FSC) criterion was introduced by Van Heel in 1986, and since then has become the standard quality measurement [77]. In Fourier space, FSC measures the normalized CC coefficient between two 3D volumes over corresponding shells \( r_i \) [78]. The FSC was derived from the Fourier ring correlation (FRC) [79].

\[
FSC_{12} = \frac{\sum_{r \in r_i} F_1(r) \cdot F_2^*(r)}{\sqrt{\sum_{r \in r_i} F_1^2(r) \cdot \sum_{r \in r_i} F_2^2(r)}} \quad (1.11)
\]

In this formula (equation 1.11), \( F(r) \) is the complex ”structure factor” at position \( r \) in Fourier space. The ”\(*\)” denotes complex conjugation. The summations are over all Fourier space voxels ”\( r \)” that are contained in the shell ”\( r_i \)”. Agreement needs to be reached about the resolution level in FSC at which the result can be regarded as reliable; we still need a cut off criterion for analyzing FSC curves. Various researchers have used fixed-value resolution thresholds in their publications, and the most fashionable is a ”0.5” value as the resolution defining threshold. Recently the value ”0.143” has been proposed as
a more realistic. However, a fixed threshold cannot account for the varying number of voxels in a Fourier shell. At present, the most widely used threshold curve in connection with 3D FSC data is the $\sigma$-factor curve [80], which states:

$$\sigma(r_i) = \frac{\sigma_{\text{factor}}}{\sqrt{n(r_i)/2}} \cdot \sqrt{n_{\text{asym}}}$$  \hspace{1cm} (1.12)

In this formula (equation 1.12), $n(r_i)$ is the number of voxels contained in a Fourier shell of radius $r_i$; the extra factor of ”2” is required because the FSC summations include all Hermitian pairs in Fourier space; $n_{\text{asym}}$ is the number of asymmetric units within the given point group symmetry (”1” for an asymmetrical object, up to ”60” for objects with icosahedral symmetry) [78]. The most frequently used value for the $\sigma$-factor threshold is 3, which indicates the choice of three standard deviations above the expected random noise fluctuations as a significance threshold.

However, using a single model for the angular assignments could result in over-refinement or overfitting, because this model’s bias towards the noise and over-optimistic low-pass filtering based on inflated resolution estimates may lead to further enhancement of the noise in the model. Consequently, during multiple refinement iterations, the amount of noise may gradually increase and final resolution estimates may be grossly exaggerated [81]. The ”gold-standard FSC [82] was proposed to solve this problem. The reasoning is that more realistic resolution estimates may be obtained by refining a separate model for two independent halves of the data, so that the two half-reconstructions have no correlation. Hence, this could make FSC curves more reliable.
1.4 Principle of electron diffraction

Although it was believed that any soluble protein that could be purified would be relatively easy to crystallize, the situation in practice is more complex. Normally, proteins do not grow into large enough crystals for X-ray crystallography, and sometimes only micro-size or nano-size crystals can be obtained. Protein crystallization has to be done in an aqueous environment. Difficulties in obtaining high-quality protein samples, as well as the sensitivity of protein samples to temperature, pH, ionic strength, and other factors, increase the complexity of the crystallization procedure [83]. Protein samples must be chemically pure and conformationally uniform [84]. Biological macromolecules have the same thermodynamic rules as inorganic compounds for supersaturation, nucleation, and crystal growth.

The interaction between the high energy electrons and atoms in the sample is much stronger than that between the X-ray photons and atoms in the sample. This is because X-rays are scattered by the electrons in an atom while electrons are scattered by the electric field of the atom. The electric field is a consequence of the combined effects of both the atom’s nuclear charge and its extra-nuclear electrons. As a consequence, micro-size or nano-size crystals can be used in electron diffraction, as the crystals can be millions of times smaller than those required for X-ray diffraction [85; 86; 87; 88; 89; 90]. The principle of diffraction of high energy electrons is similar to that of X-rays. Electrons can also be scattered in two ways: elastic or inelastic. The former, which has fairly wide angular distribution, involves no energy transfer and gives rise to high-resolution information; the latter, where in the latter method the electron loses energy to the sample has a narrow angular distribution and produces an undesired background.

The incoming plane electron wave interacts with atoms in the crystal, and secondary waves are generated, which interfere with each other (Fig. 1.9) [91]. This occurs either constructively or destructively.
Constructive interference of the electrons can be observed as spots on a fluorescent screen or recorded on film, image plates, CCD cameras or DED cameras. Laue and Bragg both explained this phenomenon, and specify where diffraction can occur.

\[ a (\cos \varphi_1 - \cos \theta_1) = h\lambda \]  \hspace{1cm} (1.13)

Here, \( h \) is any integer, \( \lambda \) is the wavelength of electrons, \( \theta_1 \) is the angle between the incident wave and the scattering plane, and \( \varphi \) is the angle between the scattered wave and the scattering plane.

This is similar for the other two basic vectors:

\[ b (\cos \varphi_2 - \cos \theta_2) = k\lambda \]  \hspace{1cm} (1.14)

\[ c (\cos \varphi_3 - \cos \theta_3) = l\lambda \]  \hspace{1cm} (1.15)

This is the Laue equation [92]. Here, \( h, k, l \) is the Miller indices,
1. Introduction

(hkl), which represent not just one plane, but the set of all identical parallel lattice planes, a family of planes. The values of (hkl) are the reciprocals of the fractions of a unit cell edge, a, b and c respectively, intersected by an appropriate plane.

Figure 1.10: The Laue equation.

Bragg’s law gives us another way of thinking about diffraction: the incident electron wave is scattered by a family of lattice plane (Fig. 1.11) [93]. If diffraction occurs, then

\[ \overrightarrow{AB} + \overrightarrow{BC} = n\lambda \] (1.16)

\[ dsin\theta + dsin\theta = n\lambda \] (1.17)

\[ 2dsin\theta = n\lambda \] (1.18)

Here, \( n \) is any integer, \( \lambda \) is the wavelength of electrons, \( d \) is the perpendicular spacing between the planes, and \( \theta \) is the angle between the incident wave and scattering plane.

The Laue equation and Bragg’s law both follow the same principle, namely that the path difference of scattered waves should be an integer.
multiple of the electron wavelength $\lambda$. Bragg’s law was actually reduced from the Laue equation.

According to Bragg’s law, the space $d$ between the diffracting planes is a function of the electron wavelength $\lambda$. The equation can be converted to another form

$$\sin \theta = \frac{1}{2(d/n)} \frac{1}{\lambda}$$

(1.19)

This means that if the scattered wave satisfies Bragg’s law, we can draw a sphere with the radius $1/\lambda$ and the diffraction spot will lie exactly on the sphere (see Fig. 1.12). This sphere is called the Ewald sphere [94]. Here, the integer number $n=1$, but it is also possible to use another integer number, such as 2 or 3. In that case, a new distance $d'$ will be generated, which corresponds to another (parallel) family of lattice planes. Additionally, the shorter the distance $d$, the larger the scattered angle $\theta$ will be.

When we discuss diffraction patterns, we are normally considering the crystal in the reciprocal space. The reciprocal space is an inverse space of real space, in which a crystal is the 3D array of lattice points with a value that is a complex number. The Ewald sphere shows the
conversion from real space to reciprocal space.

Figure 1.12: The Ewald sphere.

Electron diffraction provides a good way of yielding high resolutions to "see" aspects of the internal structure of the crystal. For each spot on the Ewald sphere, its complex-valued structure factor $F_{hkl}$ is the wave number of contributions from each volume element of electron density in the unit cell projected onto a line. The electron density of a volume element centered at $(x,y,z)$ is, roughly, the average value of $\rho(x,y,z)$ in that region. The smaller we make our volume elements, the more precisely these averages approach the correct values of $\rho(x,y,z)$ at all points. We can, in effect, make our volume elements infinitesimally small, and make the average values of $\rho(x,y,z)$ precisely equal to the actual values at every point, by integrating the function $\rho(x,y,z)$ rather than summing average values.

$$F_{hkl} = \iiint \rho(x, y, z) e^{2\pi i (hx + ky + lz)} \, dx dy dz \quad (1.20)$$

However, there is a phase problem, as the diffraction data only allow us to measure the amplitude information $|F_{hkl}|$. The amplitude of $F_{hkl}$ is proportional to the square root of the measured reflection intensities $I_{hkl}$, so structure factor amplitudes are directly obtainable from mea-
sured reflection intensities. Diffraction is a Fourier transformation in essence, therefore we can recover or simulate the electron density map by inverse Fourier transform (simulated objective lens) only after we have obtained the phase information $\alpha_{hkl}$ [95].

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}| e^{i\alpha_{hkl}} e^{-2\pi i (hx + ky + lz)}$$ (1.21)

1.5 The phase problem in 3D crystals

The phase problem is the fundamental problem in crystallography, because the phase information in diffraction patterns is lost. Fourier series without phases cannot be used for recovering the internal conformation in proteins. There are some indirect methods for estimating them, such as the molecular replacement [96; 97], where similar, previously solved structures are used as the initial phase source, and then refine iteratively until the error is acceptable. In the case of X-ray crystallography, it is actually impossible to obtain information directly; however, it is possible in the electron crystallography, because electrons can be focused by the magnetic lenses and then used to form images. It is therefore feasible to obtain the phase directly from the images in which the phase was preserved, as was done in 2D protein crystallography.

However, 3D electron crystallography is problematic, because diffraction patterns are discrete in all directions. The difference between 2D protein crystals and 3D protein crystals means it is impossible to apply the data processing in 2D protein crystallography for 3D protein crystals. Having more than one unit cell thickness in the third axis makes it impossible for 2D crystallography software to determine the orientation when tilting crystals, which means that these data cannot be merged. Scientists have therefore tried to collect diffraction and image data from a single 3D crystal by using extremely low-dose radiation and direct elec-
tron detectors \cite{28, 29, 98}. This simplifies the processing of diffraction patterns collected from a single 3D crystal, compared with 2D protein crystallography, which needs to index and merge data from hundreds of crystals. It also reduces warping, bending, and in-plane rotation because of increased crystal contacts in the 3D crystal, which helps to achieve high resolution. However, the thickness of the 3D crystals needs to be controlled, because otherwise it causes more dynamical scattering and a larger difference in defocus at the top and bottom of the crystal. With the development of microscopes, automated data collection, new detectors and better software, most of these problems have been solved. Recently, the structures of lysozyme and catalase were solved from a single 3D micro-crystal using molecular replacement \cite{29, 30, 31}. In the area of X-ray crystallography, Ryan et al. succeeded in phasing X-ray diffraction data from a low resolution single particle map \cite{99}, which means that X-ray diffraction patterns were phased from the images, while still using the molecular replacement method. On the basis of current progress, I believe that phasing the electron diffraction patterns of 3D crystals is possible, because the phase preserved in the image can be corrected accurately and can form the initial phase for the electron diffraction patterns. We have therefore tried to reconstruct a map of a lysozyme nano-crystal using the cryo-EM images as the source of phase.

1.6 Outlines of this thesis

Chapter 2 discusses how we obtained images of 3D protein nano-crystals with 300 kV cryo-EM and a Falcon direct electron detector, using flash-cooled 3D crystals of the small protein lysozyme with a thickness of 100 nm. The images were taken close to focus and were devoid of contrast to the human eye. Fourier transforms of the images revealed the reciprocal lattice up to 3 Å resolution. Chapter 3 describes how to improve the SNR and contrast of these images: it assumes that the signal is corrupted by additive noise, and the noise is assumed to be a zero-mean, stationary process. We first estimated the power spectrum amount of noise and then
suppressed it. In Chapters 4 and 5, we attempted to phase the electron diffraction data by correlating experimental intensities with intensities derived from a molecular replacement structure of which we have the atomic coordinates. The known crystal structure of lysozyme was used to test the method, and the 3D structure of a peptide crystal relevant for understanding the underlying molecular processes of Alzheimer’s disease was reconstructed from their averaged projections. Euler angles for the individual averaged projections were found in Fourier space, and the translation vector was calculated using the projection matching technique. Our work thus pilots a novel way of phasing 3D protein crystals.
1. Introduction
Chapter 2

Imaging protein three-dimensional nano-crystals with cryo-EM

Igor Nederlof, Yao-Wang Li, Marin van Heel, Jan Pieter Abrahams

1Published in Acta crystallographica. Section D, Biological crystallography, 2013.
2. Imaging protein three-dimensional nano-crystals

2.1 Abstract

Flash-cooled three-dimensional crystals of the small protein lysozyme with a thickness of the order of 100 nm were imaged by 300 kV cryo-EM on a Falcon direct electron detector. The images were taken close to focus and to the eye appeared devoid of contrast. Fourier transforms of the images revealed the reciprocal lattice up to 3 Å resolution in favourable cases and up to 4 Å resolution for about half the crystals. The reciprocal-lattice spots showed structure, indicating that the ordering of the crystals was not uniform. Data processing revealed details at higher than 2 Å resolution and indicated the presence of multiple mosaic blocks within the crystal which could be separately processed. The prospects for full three-dimensional structure determination by electron imaging of protein.

2.2 My contribution

My main contribution to this chapter was in the area of data processing. The cryo-EM images of lysozyme had poor contrast, caused by radiation damage, low electron dose, weak phase object approximation, and so on. The contrast therefore needed to be enhanced. Averaging the patches with the same orientation in the image was expected to work. However, the crystal is probably not perfect, for instance because of beam-induced motion or because it is twinned, warped or contains mosaic blocks. All of these result in local differences in orientation. Patches were therefore extracted from each image and analyzed by MSA, in order to classify them into different domains based on their orientations. The traditional image averaging method was then used to average out the noise. After this, high contrast averaged images were obtained.

2.3 Introduction

Three-dimensional protein crystals that are smaller than about 1 mm are beyond the scope of the usual diffraction methods in structural biology.
Since about 30% of proteins that crystallize do not produce crystals of a sufficient size or quality for X-ray structure determination [100; 101], this is a serious bottleneck. Structural information on important drug targets, such as membrane proteins and large complexes, is often lacking owing to the inability to grow sufficiently sized and ordered crystals. Current trends in X-ray crystallography focus on data collection from ever smaller crystals. For example, micro-focused X-ray beams and improved quantum area detectors, such as the PILATUS, have decreased the size limits on crystals [102; 103]. In particular, free-electron lasers expand the crystallographic method towards smaller crystals [104]. We believe that electron microscopy could have a large impact on the field of protein nano-crystallography since electrons are several orders of magnitude less damaging to protein crystals than X-rays per diffracted quantum [105].

Since the development of cryo-electron microscopy (cryo-EM) and macromolecular reconstructions in the 1970s [106; 107; 108], there have been constant improvements in the maximum resolution that can be achieved using this method. Recent advances in ‘single-particle analysis’ make the solution of large molecular complexes at atomic resolution imminent [40; 109]. Important recent improvements in single-particle analysis have been automated sample handling (FEI EPU: http://investor.fei.com/releasedetail.cfm?ReleaseID=495245; [110]), automated image processing [63; 67] and direct electron detection (FEI Falcon: http://investor.fei.com/releasedetail.cfm?ReleaseID=399045; [111]). Automated data collection makes it possible to collect millions of images and software can automatically average the signal and perform angular reconstructions of the protein models. Beam damage can be minimized by using efficient direct electron detectors which significantly improve the signal-to-noise ratio at low electron-dose conditions.

In material sciences, electron diffraction is a well developed tool for
2. Imaging protein three-dimensional nano-crystals

Structure determination of inorganic crystals, in which selected area electron diffraction from different zone axes can be used to determine three-dimensional unit cells. Convergent-beam electron diffraction [112] can be used to obtain more information about the symmetry of the crystal. The development of precession electron diffraction [113; 114], automated diffraction tomography [22; 115] and rotation electron diffraction [116] have contributed to the toolbox of electron crystallographers. Applications for electron crystallography of small molecules are becoming robust, making three-dimensional structure determination by electron diffraction of three-dimensional nano-crystals a very attractive method.

Structure determination using electron diffraction of two-dimensional protein crystals has been used since the seminal work on bacteriorhodopsin in the 1970s [117]. In 2005, the structure of two-dimensional aquaporin crystals was solved to a resolution of 1.9 Å [14]. However, three-dimensional nano-crystals of proteins have so far resisted structure determination. The main reasons for this are the beam sensitivity, the large unit cell and the thickness of the crystals. The latter factor contributes to increase multiple scattering and nonlinear effects in electron diffraction and imaging. However, multi-slice least-squares methods [21] have tackled such problems in electron crystallography of small molecules. To increase the resolution and refine structural models, electron-diffraction data can be combined with electron-microscopy images, which contain phase information [13; 118; 119; 120]. Phases from these electron micrograms can also be used to solve structures from X-ray diffraction where phases are missing [96; 121].

Here, we report the electron imaging of three-dimensional protein nano-crystals that were prepared using standard protein-crystallization techniques. We have previously collected rotation electron-diffraction data of similar lysozyme nano-crystals to a resolution of 1.8 Å [98]. We discuss the preliminary image processing results and the possibility of
integrating diffraction data with imaging data.

2.4 Materials and methods

2.4.1 Crystallization

Crystallization experiments were carried out using the standard sitting-drop vapour diffusion technique in Innova-dyne SD-2 plates. The Rock-Maker software (Formulatrix) was used to design the experiments. We used a Genesis (Tecan) to dispense the screening solutions into the reservoirs. An Oryx 6 crystallization robot (Douglas Instruments) was used to transfer 500 nl reservoir solution and 500 nl protein solution in sitting drop wells. Plates were stored at 298 K and imaged using the automated imaging system Rock Imager (Formulatrix). Lysozyme (8 mg/ml) formed needle crystals after 48h when mixed 1:1 with well solution containing 0.1 M Sodium Acetate pH 3.8 and 1.0 M Potassium Nitrate(Fig.2.1).

2.4.2 Vitrification

Protein crystals were vitrified using a Vitrobot (FEI: http://investor.fei.com/releasedetail.cfm?ReleaseID=255249; [122]). 3 ml well solution was mixed with the drop containing nano-crystals and transferred onto a 3 mm holey carbon grid (Agar). Excess liquid was blotted away (blot time 3 s; blot force 5) and the sample was plunge frozen in liquid ethane. Samples were transferred into the microscope immediately and loaded and kept at 93 K using an automated cryo-loader/cryo-stage.

2.4.3 Electron-microscopy data collection

Electron images were obtained using a Titan Krios (FEI) transmission electron microscope at NeCEN. The FEG was operated at 300 keV and a Falcon (FEI) 4k×4k direct electron detector was used at 0.822 and 1.055 Å per pixel. The electron dose used was between 5 and 10 e⁻/Å².
Figure 2.1: Lysozyme nano-crystals in a crystallization drop imaged with a light microscope. The needle-shaped crystals that are visible under the light microscope are too thick for EM analysis because of absorption. However, the crystallization drops also contain (much) smaller crystals with a thickness of the order of 100 nm, which could be imaged at high resolution by cryo-EM.
2.5 Results

Image data were collected from about 200 lysozyme three-dimensional nano-crystals in random orientations and at tilt angles varying between -45° and +45°. All images were collected close to the Scherzer focus (which is ∼70 nm at 300 kV) and were devoid of amplitude contrast (Fig. 2.2). However, Fourier transformation of the images indicated crystalline order (Fig. 2.3), with discernable Bragg spots extending to about 4 Å or better for about half of the crystals and even further in favourable cases (Fig. 2.4). The crystals that yielded the best resolution on average were about 100 nm thick, as determined by tomography (Fig. 2.5). We did not orient the crystals prior to high resolution imaging, so most of their Fourier transforms showed multiple Laue zones (Fig. 2.3). These multiple Laue zones appeared as two-dimensional lattices of Bragg spots with three primitive spacings; the third spacing corresponds to the distance between lunes that are observed in diffraction patterns. This is different from two-dimensional electron crystallography, as diffraction patterns of two-dimensional protein crystals only have two primitive spacings. The presence of three primitive spacings predictably leads to moiré patterns in the real-space image. The repeat of this complex pattern of fringes that is caused by the multiple Laue zones is determined by the smallest common denominator between the primitive spacings.

2.5.1 Structure of the Bragg spots

Detailed analysis revealed the Bragg peaks to be structured (Fig. 2.3). In diffraction studies, the shape (but not the intensity) of a Bragg peak is the product of the shape of the source and the variation of the crystal spacing corresponding to the index of the Bragg peak within the diffracting crystal. As we did not measure the diffraction pattern, but instead calculated the Fourier transform of an EM image, the situation was slightly different. The shape of the source turned out to be irrelevant. The observed structure of the Bragg peaks must therefore at least
2. Imaging protein three-dimensional nano-crystals

**Figure 2.2:** The shape of the crystal can be showed after calculating the local variance and its lattice contrast can be enhanced by a Wiener type filter. (a) An electron micrograph of a lysozyme three-dimensional nano-crystal collected using a Titan cryo-EM with a Falcon camera (4096 × 4096 pixels, 0.9 Å per pixel). The location of the enlargement shown in (b) is indicated. (b) Enlargement (1024×1024 pixels) of (a). Although appearing to be just noise, the image shown here in fact does contain very significant high-resolution detail, as is revealed by its Fourier transform (Fig.2.3). Calculating the local variance of the lattice-enhanced image within a circular area with a diameter of 30 pixels reveals the shape of the crystal (c). A Wiener filter can be used to enhance the lattice contrast, as shown in (d), which is equivalent to (b) after filtering.
in part be caused by the shape and internal ordering of the crystal. In some cases splitting of the diffraction spots was observed (e.g. the left most spot in the bottom part of Fig. 2.3), but splitting was by no means the norm. Like the other irregularities in the Bragg spots, splitting could be caused by non-uniform ordering of the crystal, but Ewald sphere curvature cannot be ruled out as a contributing factor.

**Figure 2.3:** Fourier transformation shows the lattice spots of crystal. Top left, slightly enlarged Fourier transform of Fig. 2.2(a). Top right, Fourier transform of Fig. 2.2, with peak positions indicated, showing a projection of a regular three-dimensional lattice (the ring represents 4.5 Å). Bottom, detail of the top left, showing the structure of the Bragg spots.

By Fourier transforming the images, we could not only calculate the
2. Imaging protein three-dimensional nano-crystals

intensity distribution within the Bragg spots, but also the phases of each reciprocal pixel. Using ImageJ [123], both the intensity and the phase of each reciprocal pixel can be inspected in a number field. Inspection showed that the phase correlation between adjacent reciprocal pixels was low within a single Bragg spot. This suggests that the observed spreading of the Bragg spots was probably caused by local differences in the size and projected potential of the unit cells (resulting in subtle origin shifts of the unit cells with respect to the average lattice), rather than by in-plane rotations of unit cells relative to one another (which would have preserved the phase).

Figure 2.4: Histogram of the maximum resolution observed in electron images of 200 different lysozyme three-dimensional nano-crystals.

2.5.2 Centrosymmetry of the Fourier transform

A detector placed in the diffraction plane of the microscope measures a diffraction pattern. Its intensities are those of the Fourier $F(h)$ of the electron-plane-wave-function $\Psi(x)$ that has passed through the sample,
\[ F(h) = F^{-1}[\Psi_e(x)] \] (2.1)

This Fourier transform of the exit wavefunction \( \Psi(x) \) peaks at the Bragg positions determined by the periodicity of the crystal. In imaging mode, the lenses of the microscope recombine these structure factors through a Fourier transform into a real-space defocused electron-plane wavefunction \( \Psi_{\Delta f}(x) \). Distortions of this recombination are described by a centrosymmetric transfer function \( \tau(h) \), which is determined by the wavelength of the electrons, aberrations of the electron lenses and the defocus (the coherence of the illumination system and the position of the detector relative to the focal imaging plane of the microscope),

\[ \Psi_{\Delta f}(x) = F[F(h)\tau(h)] \] (2.2)

Note that \( \Psi_{\Delta f}(x) \) is a complex wavefunction. When \( \Psi_{\Delta f}(x) \) impings on the detector, the wavefunction collapses and its intensities [which are calculated by multiplying \( \Psi_{\Delta f}(x) \) by its complex conjugate \( \overline{\Psi_{\Delta f}(x)} \)] are measured as the electron image \( M(x) \),

\[ M(x) = \Psi_{\Delta f}(x)\overline{\Psi_{\Delta f}(x)} \] (2.3)

Note that \( M(x) \) is a real function, so its Fourier transform \( F^{-1}[M(x)] \) must obey Hermitian symmetry: \( F^{-1}[M(x)] = \overline{F^{-1}[M(-x)]} \). However, in the general case \( F(h) \) is not centrosymmetric. Hence, the Fourier transform of the measured image cannot be equivalent to \( F(h)\tau(h) \). An example of a typical electron-diffraction pattern of a non-oriented lysozyme three-dimensional nano-crystal is shown in Fig. 2.6. The curvature of the Ewald sphere, in combination with the parallel electron beam, causes the Bragg spots to be limited to the so-called lunes that
2. Imaging protein three-dimensional nano-crystals

are such a usual and prominent feature in X-ray diffraction patterns of non-oriented protein crystals. These lunes clearly break centrosymmetry.

Nevertheless, even though the crystallographic phases corrected by \( \tau(h) \) are not necessarily always equivalent to the phases of \( \mathbf{F}^{-1}[M(x)] \), it is possible (through iterative procedures) to retrieve crystallographic phase information from the inverse Fourier transform of \( M(x) \). This is relatively straightforward if the weak phase object approximation holds. Firstly, we have to realise that the Fourier transform of \( M(x) \) is equivalent to the autocorrelation of the Fourier transform of \( \Psi_{\Delta f}(x) \). This follows from the convolution theorem,

\[
\mathbf{F}^{-1}[M(x)] = \int F_{\Delta f}(h_t)F_{\Delta f}(h+h_t)dh_t \tag{2.4}
\]

The weak phase object approximation implies that most of the \( F_{\Delta f}(h) \) is located at \( h = 0 \). This implies that we can use the following approximation for weak phase objects,

\[
F_{\Delta f}(h_t)F_{\Delta f}(h+h_t) \neq 0 \Rightarrow h_t = 0 \lor h + h_t = 0 \tag{2.5}
\]

Therefore, the product \( F_{\Delta f}(h_t)F_{\Delta f}(h+h_t) \) is only nonzero if either \( h_t = 0 \) and/or if \( h + h_t = 0 \). Using this restriction, the integral in lemma(2.4) can be simplified, as \( h_t \) can be substituted by \( 0 \) or \( -h \),

\[
\mathbf{F}^{-1}[M(x)] = \int F_{\Delta f}(0)F_{\Delta f}(h) + F_{\Delta f}(-h)F_{\Delta f}(0)dh_t \tag{2.6}
\]

This reduces to
\[
\frac{\mathbf{F}^{-1}[M(\mathbf{x})]}{F_{\triangle f}(0)} = F_{\triangle f}(\mathbf{h}) + \overline{F_{\triangle f}(-\mathbf{h})}
\] (2.7)

In other words, provided that the weak phase object approximation holds, the reverse Fourier transform of the measured image is the same as the defocused set of structure factors of the crystal (in any orientation) added to the complex complement of the same set after a point inversion through the origin (which is equivalent to a rotation by) followed by taking the complex complement.

The implications of this conjecture are as follows.

1. The reverse Fourier transform of the measured image has Friedel symmetry (which is another way of saying that the measured image is real).

2. The amplitude of a structure factor is affected provided that the diffraction pattern (rather than the inverse Fourier transform of the image) does not have another structure factor located at the location after a rotation about the centre by \(\pi\).

3. If the diffraction pattern records two Friedel mates simultaneously, the amplitude of these reflections in the inverse transform of the measured image is changed, but the phase is not affected.

4. If the diffraction pattern records two Friedel mates simultaneously but only partially (owing to the presence of mosaic blocks, for instance), splitting can occur if the two reflections are not recorded on spot-on centrosymmetric locations.

5. If the diffraction pattern records two reflections that are not Friedel mates but that occur by chance at each other’s centrosymmetric locations, the reverse transform of the image mixes the two by
2. Imaging protein three-dimensional nano-crystals

adding one to the complex complement of the other, thus imposing
Hermitian symmetry on the reverse transform.

6. Because of the imposition of centrosymmetry on the reverse trans-
form, information on the handedness is lost.

Similar considerations, albeit in different formalisms, have been dis-
cussed in papers on the implications of the curvature of the Ewald sphere
on single-particle reconstructions [124] and in two-dimensional crystallog-
raphy [125]. There are also parallels with the single-sideband approach
in electron microscopy [126].

![Figure 2.5: Two defocused (-8.00 mm) frames from raw data of a tomographic
series of one of the crystals from which the high-resolution data were collected.
Frames were taken at 0° (left) and at 44° (right) and show that the crystal was
less than 150 nm thick.](image)

2.5.3 Visualizing the lattice

The translational symmetry of an EM image of a (three-dimensional)
nano-crystal can be enhanced by first Fourier transforming the image,
then zeroing all reciprocal pixels that do not belong to the lattice and
finally reversing the Fourier transform. This procedure has been applied with great success in two-dimensional protein crystallography (see, for example, [13]). We can enhance this procedure by assuming that the signal of the crystal lattice is not correlated with the noise. This implies a phase difference of $\pi/2$ between the expected structure factors of the lattice on the one hand and noise on the other. This implies

$$|F_m(h)|^2 = |\langle F_l(h) \rangle|^2 + |\langle F_n(h) \rangle|^2$$  \hfill (2.8)

The phase of $F_m(h)$ is the best (and only) estimate of the phase of $F_l(h)$. We therefore have to project $F_l(h)$ onto $F_m(h)$ to obtain the best estimate of the structure factor corresponding to the lattice, denoted here by $F_{w,l}(h)$. The right-angled geometry that imposed lemma (2.8) also implies

$$F_{w,l}(h) = F_m(h) \frac{|\langle F_l(h) \rangle|^2}{|F_m(h)|^2}$$  \hfill (2.9)

Substitution with (2.8) results in

$$F_{w,l}(h) = F_m(h) \left[ 1 - \frac{|F_n(h)|^2}{|F_m(h)|^2} \right]$$  \hfill (2.10)

This is equivalent to an optimal (Wiener) filter, which requires calculation of the ratio between an estimate of the power spectrum of the noise $|F_n(h)|^2$ and the smoothed power spectrum of the measured image.

$^1$For each reciprocal pixel $h$, we define $F_m(h)$ to be the measured structure factor and $|F_m(h)|$ is its amplitude; $|\langle F_l(h) \rangle|$ is the amplitude of the expected structure factor of the lattice and $|\langle F_n(h) \rangle|$ is the amplitude of the expected structure factor of the noise.
2. Imaging protein three-dimensional nano-crystals

$|F_m(h)|^2$ [127]. We generated this optimal lattice filter by first smoothing the power spectrum of the image using a soft-edged circular window with a radius of 256 reciprocal pixels. We estimated the power spectrum of the noise to be the rotational average of this smoothed power spectrum. In order to prevent $F_m(h)$ being multiplied by negative numbers in (10), we applied a threshold limiting the range of its scale factor to $(0, 1)$. The application of (10) clearly revealed the lattice (Fig.2.2c). The outline of the crystal now becomes apparent on calculating the local variance of the lattice-enhanced image (Fig.2.2d).

![Figure 2.6: Electron diffraction pattern showing lunes (at 200 keV).](image)

Further image processing with IMAGIC [128] enhanced the resolution. We randomly picked 101 patches of $256 \times 256$ pixels from the lattice-enhanced image, aligned them and classified them [129] into three classes (Fig.2.7, top panel). We used these classes to identify 4500 equivalent patches in the lattice-enhanced image by a correlation search. The locations with high correlation were in agreement with the contour of the nano-crystal (compare Fig.2.2d with the top right panel of Fig.
We classified the selected patches into five different classes of on average 900 patches. Four of the class averages are shown in the second panel from the top in Fig. 2.7. A fifth class that had less pronounced contrast and was located at the edges of the crystal is not shown.

The resolution of the patch averages is illustrated using their Fourier transforms (the second panel from the bottom in Fig. 2.7). At lower contour levels, spots can be seen extending almost to the Nyquist frequency (1.64 Å). The classes differ both in real space and in reciprocal space. In reciprocal space, Bragg spots that are prominent in one amplitude spectrum are hardly visible in the other amplitude spectra and vice versa. These differences in reciprocal space are not resolution-dependent. Notice furthermore that the differences between classes 1 and 2 and between classes 3 and 4 are subtle, but that the differences between classes 1 and 4 are more pronounced.

We analyzed all crystals from which we collected images that showed Bragg spots beyond 4Å resolution. All resulted in high-resolution, high-contrast classes, some of which are shown in Fig. 2.8.

## 2.6 Discussion and conclusions

We have shown that three-dimensional protein nano-crystals that are too small for conventional (synchrotron) X-ray analysis can be robustly imaged at high resolution with electrons. Fourier transforms of these images reveal the presence of moiré lattices caused by interference between the lattices of the zero-order Laue zone with higher order Laue zones, indicating that the crystals are not aligned with their principal axes parallel to the electron beam. Furthermore, the Fourier transforms have Hermitian symmetry, which has some repercussions for extracting crystallographic phase information. As long as the weak phase object approximation is valid (which is the case when most
2. Imaging protein three-dimensional nano-crystals

Figure 2.7: The patches in the same domain can be averaged to enhance the contrast. The top row of images shows the reconstructed projection classes of the electron image of the three-dimensional nano-crystal of lysozyme shown in Fig. 2.2 using just 101 non-overlapping sub-images (patches) of 256×256 pixels randomly selected from the Wiener-filtered image shown in Fig. 2.2. Using these three initial projection classes, we selected high-correlation patches; their locations are shown on the top right. These patches were classified into five classes, four of which are shown in the second panel from the top. The patches that contributed to each of these classes are located at the positions plotted in the bottom row. Patches were cut out of the original non-filtered image (Fig. 2.2a) at exactly the same locations and were assigned their corresponding shift and class parameters. The second panel from the bottom shows the amplitudes of the Fourier transforms of their class averages at low contour level and suggests a resolution of beyond 2 Å (the resolution of the circle is 2.1 Å and there are lattice spots well beyond this circle).
Figure 2.8: Examples of averaged images with high contrast. A dozen of the more than 200 high-resolution projection images (resolution better than 2.5 Å judging from their Fourier transforms) of different lysozyme three-dimensional nano-crystals, all in different orientations. The data were collected as in Fig. 2.2 and processed as described in the text.
2. Imaging protein three-dimensional nano-crystals

electrons scatter not more than once within the sample), extracting phase information from the Fourier transform of the measured images is relatively straightforward.

The lattice can be enhanced by Wiener filtering. Knowledge of the lattice parameters is not required for this procedure. Hence, the outcome is not biased by any imposed lattice parameters, which is an advantage.

Using software designed for single-particle cryo-EM data analysis, the contrast and resolution of the projection images can be dramatically improved by identifying coherent domains of the crystal and averaging within them. With a similar aim, images of two-dimensional crystals have been analyzed in the past using multivariate statistical analysis (MSA) in order to analyze the differences between adjacent unit cells in the two-dimensional lattice [74; 130; 131]. However, here we analyzed three-dimensional nano-crystals, which some distinguishing differences compared with two-dimensional crystals. Since the third cell axis of the crystals is usually not aligned with the electron beam, the projection images usually show moiré patterns (see Fig. 2.2), which manifest themselves in the Fourier transform as multiple reciprocal two-dimensional lattices (Fig. 2.3 and 2.7). Note that the moiré unit repeat in the electron projection is larger than the projection of a single three-dimensional unit cell.

When crystals are also ordered in the third dimension they are less likely to be affected by warping and bending, or by in-plane rotations of the unit cell, because of increased crystal contacts. This will increase the probability of achieving high resolution. The increased thickness of the crystals will cause more pronounced dynamical scattering and a different defocus at the top and bottom of the crystal.

Detailed analysis revealed that within a crystal, we could observe
different views of projected density within a single crystal, correlating with the location within the image (cf. Fig. 2.7). Potential causes are the following.

1. The presence of mosaic blocks within the crystal, each with a slightly different orientation of the unit cells, resulting in different Bragg spots to be sliced by the Ewald sphere.

2. The crystal is not normal to the electron beam, resulting in a different focus for different parts of the crystal.

3. The crystal is not uniformly thick, resulting in a higher degree of dynamical scattering in the thicker parts of the crystal compared with the thinner parts.

Differential defocus is a less likely explanation of the observed differences, as this reduces the intensity of the Bragg spots in the reverse Fourier transforms uniformly within a resolution shell (provided that the microscope is stigmatically aligned). Fig. 2.7 does not suggest this to be the case. Differences in thickness of the crystal would result in a (subtle) density gradient of the image owing to increased electron absorption by thicker parts of the crystal. We did not observe such gradients. Furthermore, increased dynamical scattering can only cause the intensities of spots to become more alike, and we observe some spots that are present in one reverse transform to be absent in the other and vice versa, which cannot be explained by dynamical scattering effects. Hence, for the moment we consider the presence of mosaic blocks to be the most likely explanation of the observed differences. Macromolecular X-ray crystallography explicitly assumes and models the presence of such mosaic blocks within crystals; here, we could be seeing them directly for the first time.

Essential steps towards real space three-dimensional nanocrystallography have been taken. Next, we aim to automate the data analysis, so that projection views of many different orientations can be straightforwardly determined. We have made progress in this
2. Imaging protein three-dimensional nano-crystals

respect, as shown in Fig. 2.7. We also aim to apply three-dimensional reconstruction techniques that are well known in the single-particle cryo-EM community to the reconstruction of the three-dimensional crystal structure, initially at low resolution. However, at the resolution that we can now achieve (Fig. 2.7) we will also have to take aspects into account that single-particle analysis has so far been able to ignore: dynamical scattering and focal differences between the top and the bottom of the sample (equivalent to considering Ewald sphere curvature).

In subsequent refinement procedures with the aim of achieving high (atomic) resolution, these aspects will also have to be modelled. We assume that the knowledge that we obtain in this venture will assist in the field of single-particle analysis for analyzing structures of noncrystalline materials to resolutions at which these aspects can no longer be ignored.
Chapter 3

Filter for processing image data of 3D protein nano-crystals

Eric van Genderen*, Yao-Wang Li*, Igor Nederlof, Jan Pieter Abrahams

1Published in Acta crystallographica. Section D, Structural biology, 2016.
* These authors contributed to this work equally.
3. Filter for processing image data

3.1 Abstract

When 300 kV cryo-EM images at Scherzer focus are acquired from 100 nm thick three-dimensional protein nanocrystals using a Falcon 2 direct electron detector, Fourier transformation can reveal the crystalline lattice to surprisingly high resolutions, even though the images themselves seem to be devoid of any contrast. Here, it is reported how this lattice information can be enhanced by means of a wave finder in combination with Wiener-type maximum-likelihood filtering. This procedure paves the way towards full three-dimensional structure determination at high resolution for protein crystals.

3.2 My contribution

In this chapter, I worked on the design of the algorithm and wrote the code. The cryo-EM images of lysozyme nano-crystals and peptide nano-crystals were devoid of contrast. The contrast of the images of lysozyme nano-crystals was enhanced using MSA, as described in the last chapter, but the contrast of the images of peptide nano-crystals was not. The reason is probably the difference in molecular weight. We therefore designed a new image filter, the lattice filter, which can remove the noise in Fourier space. The noise fraction was estimated on the basis of its power spectrum and applied to its structure factor, retaining the signal fraction. Then, a contrast enhanced image was obtained by the inverse Fourier transformation. This filter is based on the hypothesis that the noise was added to the signal. Hence, the radial average of the power spectrum was selected as the noise estimation. In addition to the radial average, the mean value and median value can also be used.

3.3 Introduction

Previously, we demonstrated that electron diffraction of protein 3D nano-crystals could yield 2Å data [28]. Recently, we collected hundreds
of high-resolution electron images at Scherzer focus of cryo-preserved, randomly orientated 3D nano-crystals of our test protein lysozyme [98]. Although appearing devoid of signal, Fourier transformation revealed crystalline order to a resolution of 4 Å or better in about 50% of cases. Crystals with a thickness of about 100 nm (corresponding to 15 to 30 unit cells) yielded data with the best quality. The resolution of the Bragg spots in the Fourier transform of the electron micrograph is the lower threshold of the crystalline order. If the crystal occupies only part of the electron image, the remainder of the image contributes nothing but noise. If the crystal is cracked, twinned, warped, or contains mosaic blocks, the resolution of the Fourier transform is reduced because the unit cells do not align perfectly. In 2D crystallography the resolution is enhanced by computationally ‘unbending’ the crystal [132; 133; 134].

Firstly, the two-dimensional lattice repeat is identified and the frequencies that do not conform to this repeat are filtered from the Fourier transform of the image by setting them to zero, thus enhancing the translationally repeating features of the image. This is equivalent to averaging the image with shifted versions of itself, whereby the magnitude and direction of the shifts are determined by the lattice parameters. This procedure will therefore average out noise that does not have translational symmetry, as it is not correlated to the signal. The image of the crystal is then subdivided into patches, which are subsequently aligned and averaged [135; 136; 137].

One of the main differences between two-dimensional crystals and three-dimensional crystals is that projection images of randomly oriented three-dimensional crystals usually show moiré patterns, rather than regular two-dimensional lattices. Hence, it is not possible in the general case to extract a repeating unit: the moiré is not usually defined as a rational sum of the other two independent lattice vectors. The moiré pattern (a potentially nonrepeating pattern that results from the superposition of multiple lattices) exists because of the three
3. Filter for processing image data

independent lattice parameters characterizing the three-dimensional crystal. Owing to Ewald sphere curvature, beam divergence and/or crystal mosaicity, repeats corresponding to each of these three translational symmetries can co-exist in one and the same image. If one of the translation operators cannot be expressed as an integer sum of the other two, a moiré pattern results. While this pattern does not directly show the crystal lattice, it does contain information on this lattice.

Here, we discuss a procedure for enhancing the moiré lattice information in the analysis of three-dimensional nanocrystals that does not require knowledge of the lattice parameters or orientation of the crystal, and even allows the lattices to be enhanced if multiple crystals are present in the image.

A high-resolution image of a (three-dimensional) crystal will have translational symmetry. Although this symmetry may not be obvious because of noise, the amplitudes of a Fourier transformation of the image will reveal the reciprocal lattice. One way of enhancing the translational symmetry in the image therefore is to identify the parameters that describe the reciprocal lattice, then to zero all reciprocal pixels that do not belong to this lattice and reverse the Fourier transform as in two-dimensional crystallography. However, this approach has a mathematical flaw, which can be understood intuitively as follows. Suppose a high background and a weak Bragg spot in the Fourier transform of the image. If all reciprocal pixels are set to zero but the Bragg pixels are kept at their original value, then this will lead to an incorrect estimation of the relative strength of this particular Bragg spot. This is equivalent to overweighting a weak Bragg spot relative to a stronger Bragg spot that is higher above the background. The Wiener filter addresses this problem in a maximum-likelihood approach. An improvement over existing methods is obtained by assuming that the power spectrum of the noise and the lattice signal are uncorrelated. This is equivalent to establishing a Wiener filter that optimally enhances the lattice.
approach does not need knowledge of the crystal lattice constants nor of
the crystal orientation or location to obtain a good result.

Assume a structure factor of the Fourier transform of the image: $F_m(h)$. It is the sum of the structure factor of the crystal lattice $F_l(h)$
and the structure factor of noise $F_n(h)$:

$$F_m(h) = F_l(h) + F_n(h) \quad (3.1)$$

Neither $F_l(h)$ nor $F_n(h)$ are known. We can only assume that they
are uncorrelated. Their expected absolute phase difference will therefore
be $\pi/2$, hence together with $F_m(h)$ they define a right-angled triangle:

$$|F_m(h)|^2 = |\langle F_l(h) \rangle|^2 + |\langle F_n(h) \rangle|^2 \quad (3.2)$$

We can infer $|\langle F_n(h) \rangle|^2$ from the power spectrum of the image and
use the result to calculate the expected amplitude of $F_l(h)$. However,
we also require its phase and the only reasonable estimate is the phase
of $F_m(h)$. So we need to project $\langle F_l(h) \rangle$ onto $F_m(h)$ to get the best
estimate of the expected lattice structure factor $F_{l,b}(h)$ (see fig.3.1).

Geometry implies the following equality, which is equivant to an op-
timal (Wiener) filter [127]:

$$F_{l,b}(h) = F_m(h) \frac{|\langle F_l(h) \rangle|^2}{|F_m(h)|^2} \quad (3.3)$$

Substituting with (3.2) gives:

$$F_{l,b}(h) = F_m(h) \left( 1 - \frac{|\langle F_n(h) \rangle|^2}{|F_m(h)|^2} \right) \quad (3.4)$$

Thus, scaling the structure factors of the original image by this like-
lihood will recover the phases of the waves as well as their amplitudes.
This theory leads to a more robust algorithm as described below.
3. Filter for processing image data

Figure 3.1: The measured structure factor $F_m(h)$ is the sum of the unknown structure factor corresponding to the lattice signal $F_l(h)$ and the unknown structure factor corresponding to the noise $F_n(h)$. On average, $F_l(h)$ and $F_n(h)$ are uncorrelated, hence the expected absolute phase difference between them is $\pi/2$. We can only infer the absolute values of the expected structure factors. We cannot infer their phases, because there are two equally valid solutions mirrored over $F_m(h)$. The best estimate for $F_l(h)$ is denoted by $F_{l,b}(h)$. It can be calculated by projecting $\langle F_l(h) \rangle$ onto $F_m(h)$.

3.4 Method

In order to prevent wrap-around artifacts, we padded the images with pixels the values of which were set to the average value of the original image. The amount of padding can be defined by the user and corresponds to the expected size of the crystalline domains. The default value (used throughout the paper) is 1/16th of the image size (corresponding to 256 pixels for Falcon 2 images). The pixels of the Fourier transform $F(h)$ of an electron image $I(x) = \mathcal{F}^{-1}(F(h))$ contain complex numbers (hence they carry phase information). First we calculated the radially averaged power spectrum $|F_n(|h|)|^2$ of the image in order to approximate the power spectrum of the noise,

$$|F_n(|h|)|^2 = \int |F(|h| \{ \cos \alpha \sin \alpha \})|^2 \partial \alpha \quad (3.5)$$

The radial average fig. 3.2 is not completely smooth because of the
contributions of the spots at certain spacings. In order to correct for this, we assumed that the radial average of the noise power spectrum is a decreasing function of $|h|$. Thus, if $F_n(h)$ increases, this must be caused by the signal of Bragg spots. In this case, we keep $F_n(h)$ constant until is decreases below this value. An example of such a radial average is shown as a linear plot in fig. 3.2; this is the result from the image in fig. 3.3(a). Then, for each pixel, we calculate its significance $s(h)$ as a normalized signal-to-noise ratio,

$$s(h) = \frac{|F(h)|^2 - |F_n(|h|)|^2}{|F(h)|^2} \quad (3.6)$$

In the absence of noise ($s(h) = 1$) and when the norm\(^1\) of the signal is equal to the norm of the noise, ($s(h) = 0$). Note that $s(h)$ can be negative owing to fluctuations in the noise level. In fact, in the absence of signal, fluctuations in the noise level will cause $s(h) = 1$ to be negative for half of the reciprocal pixels! We consider a pixel to

\(^1\)The real valued norm of a complex value is the square of its absolute value: $|a + bi|^2 = a^2 + b^2$
3. Filter for processing image data

contribute significant information about the lattice when \( s(\mathbf{h}) \) is higher than a specified cutoff value. Pixels lower than the cutoff value are set to zero. As a default, we used a cutoff value \((c = 0.0)\) for all the examples in this paper. Thus, for all examples given, all pixels of the lattice filter \( L(\mathbf{h}) \) which had a norm below the radially averaged norm of \( F(\mathbf{h}) \), were set to zero. The value of the remaining pixels of the lattice filter \( L(\mathbf{h}) \) were set to:

\[
L(\mathbf{h}) = \frac{s(\mathbf{h}) - c}{1 - c} \quad (3.7)
\]

This lattice filter \( L(\mathbf{h}) \) can be still be noisy, especially if the signal is low, so we included the option of only considering pixels that are likely to belong to a Bragg spot. The lattice parameters of the image are usually unknown at this stage of data analysis. Thus, to identify potential Bragg spots, we used a method that does not require lattice parameters. Firstly, we selected 3×3 clusters of pixels in which each of the pixels had a norm that is above a specified acceptance level \( a \). We only allowed pixels to have nonzero values if they are less then a specified distance \( r \) away from any pixels within clusters which represent spots. This cutoff distance \( r \) is proportional to the reciprocal-space equivalent of the expected size of the crystalline domains. All pixels of \( L(\mathbf{h}) \) that are further away than \( r \) pixels from such a cluster of significant pixels [for which \( L(\mathbf{h}) > a \)] were set to zero. As defaults, we used an acceptance value of \( a = 0.4 \) and a Bragg spot radius of \( r = 4 \) pixels for all of the examples in this paper. This distance criterion can be suppressed by setting \( a = c \).

In addition to the lattice, this procedure also enhances other repetitive features of the image. Detector artifacts can be a major source of such spurious features. Detector artifacts in particular can be a major source of such spurious features. We found these artifacts to produce high-resolution features. We therefore included an option to filter out
such artificial signals by setting $L(h)$ to zero for all $(|h| > n|h|_{\text{max}})$. As a default value we used a Nyquist cutoff of ($n = \frac{2}{3}$) for all examples in this paper.

After having constructed the lattice filter $L(h)$ according to the procedure outlined above, we calculated the filtered image $I_f(x)$:

$$I_f(x) = \hat{\mathcal{F}}^{-1}(L(h)F(h))$$  \hspace{1cm} (3.8)

The nano-protein crystal images used here as illustration of the method, were acquired on a FEI titan Krios electron microscope at Scherzer focus of crystals with a thickness of approximately 100 nm. The data was collected with a Falcon 2 FEI camera on $4096 \times 4096$ pixels at 0.5 sec exposure time. The mean dose of the exposures was 3 to 10 $e^-/\AA^2$(for further details, see [98]).

### 3.5 Results

Three example images (see fig. 3.3) from lysozyme nano crystals show the merits of our new algorithm.

The processed images show moiré patterns typical for non-oriented 3D crystalline structures. Due to truncation errors, also some spurious repeating features will be visible in areas of the image where no crystal is present, but here the amplitudes are much lower than in the crystal. Thus the processed images will give clear indication where the crystal might be located. The moiré pattern (a potentially non-repeating pattern that results from the superposition of multiple lattices) exists because of the three independent lattice parameters characterizing the 3D crystal. Due to Ewald sphere curvature, beam divergence and/or crystal mosaicity, repeats corresponding to each of these three translational symmetries can co-exist in the one and the same image. If one of the translation operators cannot be expressed as an integer sum of the other two, a moiré pattern results. While this pattern does
Figure 3.3: Images of 100 nm thick 3D crystals, produced by a Titan Krios 300 kV FEG transmission EM, captured with a Falcon 2 camera (4096×4096 pixels) using an 0.5 sec exposure time and an illumination of 3 to 10 $e^-/\text{Å}^2$. Left: original unprocessed images; right: processed images. Top: lysozyme crystal; middle: Lysozyme crystal with several ice crystals included (shown in detail); bottom: crystal of a cross-$\beta$ peptide. Notice that in the cut-out of the middle image, we can distinguish an ice crystal (with a small unit cell and dominant spacings at $\pm 3.8$ Å at the left side of the cut-out) from the protein crystal (right side of the cut-out).
not show directly the crystal lattice it does contain information of this a lattice. The result of the filter algorithm is shown in figure 3.3. This information can then be used for further analysis as described in the paper of Nederlof Li et al. [98]

If the images are not padded as described in the Methods section, a wrap-around effect will occur and the information tends to bleed over the edge of the image into the opposite side of the image. This can be circumvented by padding the image, but when this padding is removed after lattice filtering, the resulting discontinuities at the image edges can lead to crosses centered on the Bragg spots in the reciprocal space, which could be undesirable for certain applications. Crosses can be prevented by choosing not to pad the images, or they can be suppressed by writing out filtered images without removing their padding (not shown).

Close inspection of the power spectrum of the image in figure 3.5 shows each of the spots to have a slightly different shape. The shapes of the spots differ when the projection of the crystal shows different domains. The latter we call mosaicity and shows that within a crystal, patches can have a slightly different orientation. While this does not change the meta structure of the lattice it will disrupt the moiré pattern and makes interpreting the structure more difficult. To put a positive twist on this: it also provides more orientations of the crystal.

How does our filter behave when applied to a image consisting of generated noise with the same median intensity and standard deviation as an image containing a crystal (values have been obtained from fig. 3.3a)? The result of such a filtering is shown in fig. 3.6. While some weak lattice structures can be seen, the amplitude of the wave structures is only slightly above the median pixel level. This shows that the filter is able to discriminate between an image consisting of random noise and a low signal-to-noise ratio image which contains a crystal lattice.

The algorithm includes the possibility to set a resolution cutoff corresponding to a factor of the Nyquist frequency. While this is a
3. Filter for processing image data

**Figure 3.4:** Rotational average which is subtracted from the power spectrum during lattice filtering. (The image is cropped and the central pixels with a radius of 4 pixels are removed as they are off the intensity scale).
**Figure 3.5:** Unique half of the centro-symmetric lattice filter of the top lysozyme crystal image. Clearly visible are the spots that make up the lattice of the real image. Because of mosaicity and crystal shape, the spots are not round and do not have the same shape.
powerful tool for removing certain detector artifacts or selecting a filter quality, misuse can introduce artifacts. Fig. 3.7 demonstrates the effect of the lattice filter when using a sub-optimal Nyquist cutoff. If it is too low, the high-resolution spots will be obliterated and therefore the high-resolution details of the crystal lattice will be lost. Choosing a value that is too high can result in a severe checkerboard striping and patterning owing to detector artifacts.

The other important filter parameter is the spot-selection threshold.

1 The artifacts introduced by a detector have structure-like features and will be of major influence in the final filtered image and should be taken into account. In the case of the Falcon 2 camera, used for the showcase sample images, this can include chip to chip fluctuations and read-out gain reference artifacts. The later are most clearly visible at the Nyquist frequency in the power spectrum of the image. The Nyquist frequency artifact reveals itself as a bright line of pixels around the edge of the power spectrum, and a checkerboard or striped patterning over the whole filtered image.
If a value is chosen above the noise level it will lead to the selection of bogus Bragg spots. This will introduce spurious noise in the filtered image. Another case is when the criteria for selecting the spots are too stringent and only a few of the brightest spots remain. In this case the final lattice image usually shows a one-dimensional or two-dimensional lattice pattern over the whole image, no longer discriminating between the crystal and its disordered surroundings.

**Figure 3.7:** Results from the same area as in figure fig:original-processed and results. (a) Image with selecting a very small area of spots far from nyquist frequency. (b) Including the nyquist frequency. Visible are the artifacts from both including the camera as well as effects from wrapping effects in the reciprocal space. (c) Low spot threshold selection; by including spots that are not Bragg spots, more noise is present, but also more detail in the lattice is visible. (d) Very high spot selection threshold, thereby selecting only the strongest spots.
3. Filter for processing image data

3.6 Discussion

In earlier work we reported interactive image processing to enhance the lattice. Although this produced good results, it was slow, tedious and required expert knowledge. Here, we show the mathematical proof of our new approach, captured in an automatic algorithm that is very fast (half a second for a 4096×4096 image on a standard 2014 desktop computer). The lattice filter is a very powerful tool for selecting and analyzing extremely low contrast cryo-images of three-dimensional protein/peptide nanocrystals. It confirms that the three-dimensional crystals are made up from multiple domains which are slightly differently oriented. Indeed, the algorithm can comfortably deal with multiple crystals with very different orientations, unit cells and/or space groups, as is witnessed in the middle panel of Fig. 3.3, which shows that the lattice of an ice crystal is enhanced just as well as the lattice of a protein crystal. Since more than two lattice parameters are required to describe the moir lattice of a projected three-dimensional crystal, approaches from two-dimensional crystallography cannot be applied straightforwardly or without considerable reprogramming. While one can argue that patches of the crystal with different orientations should not be back-transformed together, it is something that can be performed directly after the filtering process [98].

Our method does not correct for the contrast-transfer function (CTF), but since it does not affect the phases of the projection image, a CTF correction can be performed after lattice filtering. Although in principle a CTF correction could precede the lattice filter, it advisable to first perform the filtering, since this also takes care of background removal. In the examples that we give here we did not perform any CTF corrections, as the data were collected at Scherzer focus, where the first sign reversal of the CTF occurs beyond the resolution limit of our images.

We propose the new lattice filter as a powerful tool for processing very
noisy images with crystal structure factors (and thus with phase information) hidden within them. The filter is able to discriminate between noise images and the very noisy images with very low contrast which contain crystal-like structures. The lattice filter retains the shape of the spots in Fourier space and also retains any phase gradients within the Bragg spots (which determine the domain structure within the crystal). Thus, it retains all of the significant information from the Bragg spots. This will open the way to combining the phases acquired from stationary, two-dimensional images with intensities of rotation diffraction data taken from the same type of crystals. In this way, we expect to be able to phase the diffraction information of protein and peptide crystals.
3. Filter for processing image data
Chapter 4

A new approach of estimating crystallographic phases using images of 3D crystals of macromolecules

Yao-Wang Li, Tim Gruene, Jan Pieter Abrahams
4. A new approach of estimating crystallographic phases

4.1 Abstract

Knowledge of the crystallographic phases is essential for solving structures with crystallography. It is relatively straightforward to determine the unit cell and the diffracted intensities from the diffraction patterns of crystals, yet it is difficult to estimate the phases. Direct methods and molecular replacement can provide initial phases, but require atomic resolution or the structure of a similar molecule. Here we describe how phases of an unknown protein nano-crystal may be estimated using cryo-electron microscopy (cryo-EM) images. For this purpose, we used a known crystal structure of lysozyme and an unknown crystal structure of peptide. This chapter shows how to determine the in-plane shift vector and orientations (5 degrees of freedom) for the cryo-EM images of a known lysozyme nano-crystal, the structure of which had previously been solved from diffraction patterns using molecular replacement. I then discuss how to reconstruct the electron density map from these cryo-EM images. Finally, I describe how to determine the orientation of cryo-EM images of peptide VQIVYK nano-crystals from a simulated three-dimensional (3D) lattice, using no structural information except the unit cell parameters.

4.2 Introduction

It is difficult to grow protein crystals. Some protein crystals of sufficient size can be used directly in X-ray crystallography, but some protein form micro-crystals or nano-crystals, which cannot be used in a straightforward way because of radiation damage [105]. Techniques have therefore been developed to improve the efficiency of data collection, such as decreasing the voltage, improving the detector, and preserving the sample in a frozen-hydrated state. Femtosecond X-ray diffraction can solve these small protein crystals, but it requires hundreds of crystals and the instruments are expensive. It therefore cannot be used widely around the world [29; 98; 104; 138; 139]. Electron microscopy is the new trend for these micro- and nano-crystals, since electrons interact
with the atoms in the sample more strongly than X-rays, and deposit less energy. This means that smaller crystals, which cannot be used in X-ray crystallography, can be used in electron crystallography [105].

Our research group succeeded in recording rotation diffraction data from a single lysozyme nano-crystal with a direct electron detector under low-dose electron radiation [28]. A method, named MicroED, was later developed to collect rotation diffraction data from a single lysozyme micro-crystal, which it solved at the 2.9 Å resolution level. MicroED thus offers a unique and widely applicable approach for protein micro-crystal [29; 140]. The crystal used in MicroED was frozen with the liquid ethane, and the data collected at 200 keV using a CMOS detector. The electron dose was kept below 0.01 e−/Å per second, and each frame of a data set was taken with an exposure time of up to 10 s per frame. 90 diffraction patterns were collected from a single lysozyme micro-crystal (0.5-1 µm thick, 1-6 µm long and wide). This means that the fast-freezing technique, low-dose radiation and the new detector make it possible to record rotation diffraction patterns from a single crystal: either a micro-crystal or a nano-crystal.

The detector used in our research group was improved further; for electron nano-diffraction [28; 141], we built a dedicated electron diffraction camera based on four Timepix-quad quantum area detectors. Diffraction patterns of thin needle-shaped crystals were measured from a cryo-cooled 3D hen egg-white lysozyme crystal, and the results showed that this crystal with a diffracted volume that was three orders of magnitude lower than in MicroED. A dose rate of 0.04 e−/Å per frame was used in our strategy. We determined the unit cell and solved the structure at a resolution of 2.4 Å. The unit cell parameters were a=103.9 Å, b=71.2 Å, c=32 Å; the space group was P2₁2₁2₁.

The phase determination in both other group and our group was still completed through molecular replacement. However, a molecular
4. A new approach of estimating crystallographic phases

replacement model is not always available, while the images of crystals can always be measured. We therefore aimed to develop a method for using these images to phase the diffraction data. It was feasible to collect cryo-EM images of crystals in low-dose radiation [98]. We collected electron imaging data of the aforementioned lysozyme nano-crystals at 300 keV, close to Scherzer focus and with a dose of about 6 e⁻/Å². These images had poor contrast and were single-shot images of crystals with unknown orientation and position. The contrast was enhanced by multivariate statistical analysis (MSA), which classified the patches of one cryo-EM image into different groups and averaged the patches in the same group to produce high-contrast averaged images. MSA does not always work on cryo-EM images, however, depending on the quality of the data. The contrast of the cryo-EM images of peptide nano-crystals was not dramatically improved, therefore, we developed a new image filter that can enhance the contrast [142].

In order to phase diffraction patterns, the images need to be reconstructed into a map. However, determining the in-plane shift vector and orientation for the crystal’s image, which is essential for phasing, is more complicated. That is because the common center is difficult to determine, as the unit cell is reproducible in 3D space. Here, I describe a method for determining the positions and orientations of electron images. A density map was created from the lysozyme PDB file of the solved structure, and then the projections were generated from this with evenly spaced sample orientation angles. These projections were used as a reference to align the electron images (experimental projections) in Fourier space and real space. The rotated and translated experimental projections were combined with the Euler angles to reconstruct a density map, which should be similar to the reference density map. Finally, cryo-EM images of peptide nano-crystals with an unknown crystal structure were processed, using a similar procedure to calculate the five parameters. This time, no structural information was available for the reference. Hence, it was not possible to calculate the in-plane shift vector directly, because the
projections of the 3D lattice do not contain structural information and therefore cannot be referenced in real space. This work is complemented with iterative alignment in real space.

4.3 Materials and methods for lysozyme

4.3.1 Cryo-EM images and diffraction patterns

A stack of cryo-EM images of lysozyme nano-crystals (86 images) was collected at spatial sampling of 0.822 Å/pixel. The averaged projections were obtained from these (5 averaged projections for each cryo-EM image), using the MSA to enhance the contrast, and the hr data set (77 averaged images) was built, with greatly enhanced contrast of images. Diffraction patterns were collected from a single crystal, and the rotation axis and unit cell parameters of the crystal were determined by the program RED [27]. The data were processed with XDS [143]. The phases were found by molecular replacement with a monomeric polyalanine model of lysozyme (ref: PDB-2YBL).

4.3.2 Forward projection

On the basis of the pdb file lysozyme.pdb created from the solved lysozyme structure, a model density map was created using the programs Amber [144] and EMAN2 [67]. This procedure of creating a theoretical density map (map-0) comprises three steps.

1. Creating a PDB file containing a complete unit cell lysozyme_uc.pdb from lysozyme.pdb. The purpose of this step is to generate all the molecules in a unit cell.

2. Creating a PDB file of a lysozyme crystal, lysozyme_crystal.pdb. A crystal was formed by stacking the unit cells in 3D space (5 unit cells in the a axis; 7 unit cells in the b axis; 15 unit cells in the c axis).
4. A new approach of estimating crystallographic phases

3. Generating the density map (map-0.hdf) from the PDB file lysozyme_crystal.pdb. The program e2pdb2mrc.py in EMAN2 creates an electron density map (map-0.hdf) using lysozyme_crystal.pdb. The resolution (–res=5) is equivalent to standard cryo-EM definition, using $\frac{1}{2}$ Gaussian width in Fourier space. Here the argument –apix indicates the voxel size in the 3D space (Å/voxel).

UnitCell -p lysozyme.pdb -o lysozyme_uc.pdb
PropPDB -p lysozyme_uc.pdb -o lysozyme_crystal.pdb
e2pdb2mrc.py lysozyme_crystal.pdb map-0.hdf –apix=0.822 –res=5

A set of reference projections was created from the map in IMAGIC [63], and we assumed the point group C1 and sampled the rotation space in steps of 2 degrees.

4.3.3 Comparison in Fourier space and real space

The orientation of each experimental projection ($\alpha$, $\beta$ and $\gamma$) was calculated in Fourier space, and its in-plane shift vector ($x$, $y$) was determined in real space through projection matching. For the orientation determination, the experimental projections and reference projections were transformed into Fourier space, in which they are centro-symmetrical. It is therefore only necessary to perform rotational alignment between them using cross correlation (CC). The CC value determines which reference projection is the best matched projection. On the basis of the matched reference projection, the orientation for the experimental projection was found. We then came back to real space, where the experimental projection was rotated inversely, based on the rotational angle determined in Fourier space. Next, it was aligned to its corresponding reference projection translationally. On the basis of the CC value, the in-plane shift vector was calculated. Some IMAGIC scripts were prepared for this computational work. The alignment in Fourier space was used in
the procedure for creating map-2 (Figure 4.2), and the alignment in real space was used in the procedure for creating map-1 (Figure 4.3).

4.3.4 Reconstruction and validation

The rotated and shifted experimental projections were combined with the Euler angles to reconstruct a map (map-1, see Figure 4.3) in the IMAGIC program. As the control group, the matched reference projections (anchor-matched) and all the reference projections were used to reconstruct map-2 and map-3 respectively (see Figures 4.2 and 4.1). The reconstructed density maps (map-1, map-2 and map-3) and map-0 were compared internally using Fourier shell correlation (FSC) [77]. In addition, map-4 was created from the experimental projections with random orientations, in order to check the effect of reconstruction with random orientation.

![Flowchart](image)

**Figure 4.1:** The flowchart of reconstructing map-3. A set of reference projections (anchor) were generated from map-0, with symmetry C1 and angular sample size 2 degrees. Next, these projections were cut (anchor-cut) to size 256 by 256, to make them consistent with the size of the experimental projections. These projections were then used to reconstruct map-3, as the reference for comparing map-0, map-1 and map-2.
4. A new approach of estimating crystallographic phases

Figure 4.2: The flowchart of creating map-2. The reference projections were shifted using the zero-float method in IMAGIC (anchor-cut-zero). Next, the anchor-cut-zero and experimental projections (hr) were processed using the Fourier transform, to calculate their respective amplitudes (anchor-cut-zero-amp, hr-amp). After this, they were aligned to determine the best hits. The corresponding reference projections (matched reference projections, anchor-matched) were picked and used to reconstruct map-2.
Figure 4.3: The flowchart of creating map-1. The rotation angles determined in Fourier space (see Figure 4.2) were used to rotate the experimental projections inversely in the data set \( hr \) (\( hr\text{-}rotated \)). Then, the images in the \( hr\text{-}rotated \) data set were shifted by calculating the shift vectors compared with the matched reference projections, to form the data set \( hr\text{-}aligned \). Finally, these experimental projections in \( hr\text{-}aligned \) were reconstructed to produce map-1.
4. A new approach of estimating crystallographic phases

4.4 Materials and methods for peptide VQIVYK

4.4.1 Cryo-EM images and diffraction patterns

Samples containing VQIVYK crystals were prepared by pipetting a 2 µL drop of liquid containing crystals onto holey carbon film (Agar) supported by a 00-mesh copper grid. The grid was blotted and immediately plunge-frozen in liquid ethane at liquid nitrogen temperature using an FEI Vitrobot Mk. Data were collected using an FEI Titan Krios 1 cryo transmission electron microscope equipped with a Falcon (FEI) 4k×4k DED, at an accelerating voltage of 300 keV. The cryo-samples were imaged at a nominal magnification of 1.055 Å/pixel using electron doses between 5 and 10 e−/Å², and 35 images were collected.

Diffraction patterns were also collected from a single peptide nanocrystal, and currently only the unit cell parameters were determined: a=6.197, b=10.803, c=29.422; α = 90.06, β = 111.92, γ = 91.33.

4.4.2 Contrast improvement

For each cryo-EM image, 5 averaged projections were calculated. The experimental projections (77 projections) were selected from these. MSA did not dramatically improve the contrast, therefore a lattice filter was developed to enhance the contrast.

4.4.2.1 Lattice filter

The lattice filter (Equation 4.1) can calculate the fraction of signals in Fourier space [145]. The signal fraction was calculated by estimating noise, based on the hypothesis that noise was added to signal. First the power spectrum was calculated from the experimental projection. Then the spots and their shapes were determined. Next, the noise was estimated by calculating the radial averages, which did not involve the
spots and their shapes. After this, the signal fraction was calculated and the clean image was obtained using the inverse Fourier transform.

\[ H^2[f] = \frac{S^2[f]}{S^2[f] + N^2[f]} = 1 - \frac{N^2[f]}{S^2[f] + N^2[f]} \] (4.1)

The lattice filter. The signal fraction, represented by \( H^2[f] \), was determined by estimating the noise \( (N^2[f]) \).

### 4.4.3 The simulated 3D lattice and forward projection

A PDB file of peptide VQIVYK (ref: 4NP8 [146]) was modified as follows: only one atom was kept, which represented the unit cell. The unit cell parameters in PDB were replaced by our own (the space group is not known; this was not necessary in this experiment, as I was only aiming to build a 3D lattice model). The remaining steps were the same as the steps for lysozyme in creating the 3D lattice and generating the projections for the alignment in Fourier space.

### 4.4.4 Comparison in Fourier space and real space

The procedure for alignment in Fourier space was the same for the peptide experimental projections, as for the lysozyme experimental projections. The determined rotation angles and corresponding orientations (Euler angles) were passed to the peptide experimental projections. However, a different method was used to search for the in-plane shift vectors, because no structural information is available for the reference projections and the experimental projections cannot be aligned directly. In real space, the rotated experimental projections were used to reconstruct an map, from which the new reference projections were generated. Then the experimental projections were aligned in parallel to the new reference projections to calculate the translation (shift vector). Finally, the shifted experimental projections were reconstructed into a new reference.
4. A new approach of estimating crystallographic phases

map for further next refinement. This is an iterative procedure, as shown below:

1. Reconstruct the map (new reference map) from cryo-EM images.

2. Create the new reference projections from the map.

3. Conduct the parallel alignment between the experimental projections and the reference projections, in which only the translation is allowed.
Chapter 5

Preliminary results of phasing crystallographic data with images

Yao-Wang Li, Tim Gruene, Jan Pieter Abrahams
5. Preliminary results of phasing crystallographic data

5.1 Abstract

In Chapter 4, I described a method for reconstructing maps from the cryo-EM images of crystals, making it possible to phase the diffraction patterns. The phase result was verified by comparing it with that of electron diffraction structure determination using molecular replacement. In this chapter, the results are described and discussed.

5.2 Introduction

The phases of crystals have been resolved using several established methods, including molecular replacement, which requires atomic resolution or the structure of a similar molecule. However, in many cases this similar molecule probably does not exist. With the development of cryo-EM and single particle analysis, images of crystals can be recorded and used to reconstruct a map, which could solve the phase problem in crystallography. We therefore developed a method that can provide phasing from images.

We attempted to align the experimental projections and reference projections in Fourier space and real space, in order to determine the five parameters ($\alpha, \beta, \gamma, x$ and $y$). These are Euler angles and the shift vector. The lysozyme cryo-EM data were aligned to the solved structure as the reference for obtaining the five parameters. In order to verify the method, the density map ($map-1$) reconstructed from these images of crystals, and the maps ($map-2$, $map-3$ and $map-0$) reconstructed from the reference projections were compared with each other using FSC [77]. In the next step, the cryo-EM data of peptide nano-crystals were processed using a similar procedure. The Euler angles were calculated with the same procedure, while the shift vector was determined through the alignment in real space between them and the reference projections of the reconstructed map.
5.3 Results and discussion for lysozyme

The shift vector and orientation have to be determined for the reconstruction by aligning the experimental projections to reference projections in Fourier and real space. As we know, the cryo-EM images of a crystal are in a plane perpendicular to the incident beam, therefore the z-position of the crystal is irrelevant. We therefore only need to take account of the remaining 5 degrees of freedom: the shift vector \((x, y)\) and the Euler angles \((\alpha, \beta\) and \(\gamma\)). In order to calculate the shift vector of these experimental projections, they must have the same orientation in two-dimensional (2D) space, and the translation is then calculated. In other words, the amplitudes of the experimental projections must be rotated to align with the amplitudes of the reference projections, and the experimental projections are shifted in real space to match the reference projections. If an experimental projection can match one of the reference projections with a high CC value, then the orientation (Euler angles) of the matched reference projection can be passed to the experimental projection directly, since they share the same orientation in 3D space.

In order to verify this method, the known lysozyme crystal was used. The experimental projections were aligned to the reference projections in Fourier and real space. With the determined positions and orientations, these experimental projections were used to reconstruct an electron density map. In addition, all the reference projections and the reference projections that matched the experimental projections were also used to reconstruct respective maps, in order to confirm that the method is valid.

5.3.1 Reference and experimental projections

The reference projections were generated from map-0. Map-0 was created from the PDB file, which is an important theoretical map for cryo-EM images of lysozyme crystals. Generally, the space group \(P2_12_12\) (point group D2) determined from diffraction patterns is useful when
5. Preliminary results of phasing crystallographic data

generating reference projections from map-0 in IMAGIC; it reduces the number of reference projections and accelerates the calculation speed. However, we assumed that it was a non-symmetric crystal (point group C1) in order to generate the greatest number of reference projections evenly on the sphere, in case a potential reference projection that could probably be matched was missed. In addition, the Euler angle sample size is important, as it affects the precision and computational time.

The experimental projections collected in cryo-EM were close to focus, and devoid of amplitude contrast. The low contrast was mainly caused by the necessity of minimizing radiation damage. In order to reconstruct the map, high contrast images must be obtained. They were therefore processed to enhance their SNR and contrast by the means of MSA and a lattice filter. MSA can classify the patches of one image into different domains based on small differences in their orientations, and then further process these patches in the same domain to average out the noise. It worked perfectly on the cryo-EM images of lysozyme nano-crystals, but did not work on the cryo-EM images of peptide nano-crystals. The contrast of the images of peptide crystals was worse than that of the images of lysozyme crystals, possibly because of their lower molecular weight. The new lattice filter was then developed for this case; it separates the signal from the noise, and thus enhances the contrast.

5.3.2 Determination of freedom for the experimental projections

Figure 5.1 shows the result of 4 experimental projections aligned with reference projections in Fourier space; for each amplitude of the Fourier transformation of the experimental projections, the CC values for the amplitude of the Fourier transformation of the reference projections were sorted from high to low, and the two highest CC values were discriminated from the remaining 10246 reference projections. The two highest CC values of the list, are close or the same, because of
the symmetry of crystal. There are therefore two or more reference projections that can give the same CC values when compared with the experimental projections. The top two rows in Figure 5.2 indicate that the main spots were matched, although the experimental projections lack high frequency information, and the amplitude of the Fourier transformation of the experimental projections and that of the matched reference projections have the same orientation in 2D space after the amplitude of the Fourier transformation of the experimental projections has been reverse-rotated using the rotation angles.

The purpose of comparing the amplitude of the Fourier transformation of the experimental projections and that of reference projections is to ensure that the experimental projections have the same orientation as the matched reference projections. In Fourier space, this decreases the complexity because the amplitudes of the Fourier transformation of images are centro-symmetric. However, I still need to take account of other factors that could affect the result, such as the low resolution. The absence of some spots in the amplitude spectrum of the experimental projections can reduce the correlation with the model. Nevertheless, if the main spots are matched and have the same orientation in 2D, I can assume that I have succeeded in finding the right orientation in 2D space, and further calculate the rotation angle.

In real space, the exact positions of the rotated experimental projections were determined. The experimental projections were rotated using the rotation angles found in Fourier space, in order to keep the same orientation as the reference projections in 2D space. Figure 5.2 indicates that they are the same orientation (the last two rows). Then the experimental projections were shifted using the matched reference projections as the respective references. Finally, the shift vectors of these experimental projections were determined. In other words, they were centered and have the common center in 3D space, because the matched reference projections are centered.
5. Preliminary results of phasing crystallographic data

In addition, the Euler angles were also fixed, because the matched reference projections define the orientations in 3D space. Hence, the experimental projections can use the orientations determined from the model based on diffraction data directly, which would allow reconstructing the electron density map.

The shift vector \((x, y)\) of each experimental projection could be determined by comparing it with the reference projections in Fourier space and real space. The result is certainly greatly affected by the quality of the cryo-EM images, such as the resolution and the SNR. It is an effective way to align the amplitude spectrum of their Fourier transformation. These amplitude spectra are centro-symmetric and have clear spots with less noise, which is helpful in comparing them. If I had rotated and shifted the untransformed images directly in real space, I would have had to consider all five orientation parameters simultaneously, which is a much more complex problem. The program I used would also have needed to consider the rotation and translation together, which was not feasible. All the disadvantages would prevent the correct shift vector to be found. It is therefore reasonable and robust to rotate in Fourier space and then shift in real space. All in all, this method could be used to determine the shift vector, and the precision and CC values could be improved if the cryo-EM images have high SNR and resolution.

5.3.3 Map validation

If the shift vector \((x, y)\) and Euler angles \((\alpha, \beta, \gamma)\) of the experimental projections were determined correctly, then the reconstructed map should also be correct. However, the resolution that can be achieved also depends on the quality of the experimental projections, the number of experimental projections, the orientation distribution, and the reconstruction method. I therefore designed some tests to verify our method and the resolution achieved. FSC was used to estimate
Figure 5.1: The CC value curves in Fourier space. These show the amplitude of the Fourier transform of the 10284 reference projections that were aligned to the amplitude of the Fourier transform of the experimental projections by rotation. This figure gives four examples (for the experimental projections 40, 37, 39, and 16) and the results were sorted by CC value from high to low. The reference projection with the highest CC value is separated from the remaining CC’s. The corresponding projections are shown in the Figure 5.2 (the first four projections).
5. Preliminary results of phasing crystallographic data

Figure 5.2: The experimental projections and reference projections were compared in Fourier space and real space. Here, only the best 5 hits are shown (the experimental projections 40, 37, 39, 16, and 31). The top row and second row are the comparison results in Fourier space. The top row shows the amplitude of the Fourier transformation of the matched reference projections, while the second row shows the amplitude of the Fourier transformation of the experimental projections. The amplitude of the Fourier transformation of the reference projections was rotated to match the amplitude of the Fourier transformation of the experimental projections (rotation only). More Bragg spots, especially in the high frequency field, can be found on the amplitude of the Fourier transformation of the reference projections than on that of the experimental projections. However, the main spots on them, which determine the unit cell, were matched. The third and bottom rows show the comparison in real space (after the translational alignment). The experimental projections were reverse-rotated using the rotation angles and aligned to the reference projections.
how similar the reconstructed maps are to the maps from the reference projections. The FSC tests were designed as follows:

*Map-0* was created from the PDB file of the structure as determined by molecular replacement from the electron diffraction data of similar crystals.

*Map-1* was reconstructed from the experimental projections.

*Map-2* was reconstructed from the matched reference projections corresponding to the experimental projections. These matched reference projections had the highest CC values compared with the other reference projections when the experimental projections were aligned to them in Fourier space.

*Map-3* was reconstructed from all the reference projections.

*Map-4* was reconstructed from the experimental projections with random orientations.

1. **Test 1.** *Map-1* versus *map-0*. *Map-0* is the reference map for *map-1*. The result of FSC therefore shows how well they matched. The FSC curve on the left of the top row in the Figure 5.3 shows that *map-1* is similar to *map-0* at the resolution 8 Å.

2. **Test 2.** *Map-2* versus *map-0*. This indicates the theoretical resolution of *map-1*. The corresponding FSC curve on the right of the top row in the Figure 5.3 illustrates that *map-1* and *map-0* display similarity at the theoretical resolution 3 Å.

3. **Test 3.** *map-3* versus *map-0*. This is to investigate the effect of the algorithm. *Map-3* should have a high similarity to *map-0* in high resolution, and this was confirmed by the FSC curve on the left of the second row in Figure 5.3.

4. **Test 4.** *Map-1* versus *map-2*. This was designed to evaluate the quality of the experimental projection. The FSC curve shown on the right of the second row in Figure 5.3 demonstrates a resolution of 8 Å. This confirms the resolution of *map-1* from another
5. Preliminary results of phasing crystallographic data

The difference comparing with the matched reference projections was caused by the low resolution. The comparison result can also be verified by the results of test 1 and test 2. The similarity between map-1 and map-0 is at 3 Å for the theoretical resolution and at 8 Å for the practical resolution.

5. Test 5. Map-2 versus map-3. The orientation is the important information for reconstruction, and includes 5 degrees of freedom. The shift vector (x,y) and the rotational angle α (around the x-axis) are in-plane information, and do not determine the relationship out-of-plane, the β angle describes how the crystal was rotated into plane in IMAGIC (around the y-axis), and the γ angle represents how the crystal was rotated around the z-axis [38]. Hence, it is important to verify how the β angle affects the reconstructed map. The purpose of comparing map-2 and map-3 is to observe the beta angle distribution in the Euler angles. Figure 5.5 indicates that the beta angles of most of the experimental projections are concentrated in a small range (80 to 100 degrees), which reduces the resolution of the reconstructed map in the z-direction.

6. Test 6. map-1 versus map-3. This test was designed to confirm the result of test 5. The FSC curve on the right of the bottom row in Figure 5.3 shows the subtle difference compared with the FSC curve on the right of the second row (test 4).

7. Test 7. map-4 versus map-0 and map-1. The purpose of this test was mainly to check the effect of orientation. As map-4 was created with random orientation, it should have no correlation with map-1 and map-0. The FSC tests verify this (see Figure 5.4): the FSC curves are below the expected random noise level.

The conclusions of the FSC curves were verified by the orthogonal views of the four maps. Figure 5.6 shows that map-3 has the highest similarity to map-0, whereas map-2 has higher similarity to map-0 than map-1. The resolution of a map is affected by many...
factors. The orientation distribution, and especially the angles $\beta$ and $\gamma$ in this, are important. The angle $\beta$ determines the orientation distribution out-of-plane. If the angle $\beta$ distribution lies within a small range, this will dramatically affect the resolution in the $z$-direction.

### 5.4 Results and discussion for VQIVYK peptide

The procedure to determine the structure using the experimental projections of peptide is similar to using the experimental projections of lysozyme. However, there is a difference: the shift vector cannot be determined by comparison with the reference projections in real space directly, because the reference model is a 3D lattice instead of a simulated electron density map: we created the reference model by inserting just a single atom in the unit cell. So the unit cell parameters of the reference and the experimental data are the same, but the contents of the model and experimental unit cells does not correlate. Nevertheless, they can still be compared in Fourier space, because the location of Bragg spots in the amplitude spectrum of the Fourier transformation of the projections of the 3D lattice is the same as in the Fourier transformation of the projections of a simulated electron density map. The experimental projections can be rotated to obtain the same orientation of the matched reference projections in 2D space and the correct orientation (Euler angles) in 3D space. Hence, the main point of this experiment is to calculate the shift vector only using a different method.

Generally, in single particle reconstruction, the shift vector of particles can be estimated by creating an initial model from the class-averaged images. I therefore used this method to estimate the shift vector: first, to reconstruct the reference model from the experimental projections;
5. Preliminary results of phasing crystallographic data

Figure 5.3: The FSC curves between the four maps (three reconstructed maps and the original map). The FSC curves reached values above an expected random noise correlation level in conjunction with the 3σ threshold curve. The left figure in the top row shows that the resolution of map-1 is around 8.1 Å. The right figure in the middle row shows that map-1 versus map-2 displays similarity at 7.2 Å. The right figure in the bottom row demonstrates that map-2 versus map-3 displays similarity at 7.3 Å.
Figure 5.4: The FSC curves of map-4 with map-0 and map-1. Map-4 was created from the experimental projections with random orientations. The FSC values (map-4 versus map-1 and map-4 versus map-0) are below the expected random noise correlation level in conjunction with the $3\sigma$ threshold curve.

Figure 5.5: The two $\beta$ distributions in the Euler angles. The $\beta$ angle determines the relationship out-of-plane, therefore its distribution is very important. The left figure shows the beta angle distribution of all the reference projections, evenly generated by the program in IMAGIC. The right figure shows the beta angle distribution of the experimental projections (the range is around 80 to 100 degrees, apart from two projections whose beta angles are close to 10 degrees.)
5. Preliminary results of phasing crystallographic data

Figure 5.6: The orthogonal views of the reconstructed maps and the original map. The three different views indicate how well they matched. The views of the top row are from map-1, the views in the second row are from map-2, the views in the third row are from map-3, and the bottom row contains the views of map-0. Map-3 has the highest similarity to map-0, then map-2, and finally is map-1.
second, to generate the reference projections with the same orientation as the experimental projections; third, to align the experimental projections to the reference projections (translation only); and fourth, to re-reconstruct the reference model using the shifted experimental projections.

5.4.1 Contrast improvement using lattice filter

The cryo-EM data were processed by using MSA, but the contrast was still poor. I therefore further processed them using a lattice filter. Figure 5.7 shows that the contrast of an averaged EM projection was enhanced.

![Figure 5.7: Noise was suppressed by the lattice filter. Top-left is an averaged EM projection, top-right is its power spectra; bottom-left is the 'clean' image, and bottom-right is its power spectra. The SNR was improved and verified in Fourier space and real space.](image)

Although the lattice filter is faster and robust in denoising, it still has
5. Preliminary results of phasing crystallographic data

some problems; for instance, high resolution frequency can be lost, and some signals below the noise level can also be removed in Fourier space. It is therefore better to use this filter on data that still have too much noise after processing by MSA to generate the improved images.

5.4.2 Comparison in Fourier space

The experimental projections of the peptide crystal were processed in the same way as the experimental projections of lysozyme. The amplitude of the Fourier transformation of the reference projections was aligned to the amplitude of the Fourier transformation of the experimental projections. The experimental projections were then rotated and the orientations obtained. Figure 5.8 shows the results of 4 examples; the amplitude of the Fourier transformation of the experimental projections was matched to the amplitude of the Fourier transformation of the reference projections in the low frequency field. In addition, for each amplitude of the Fourier transformation of the experimental projections, the CC values of the amplitude of the Fourier transformation of the reference projections were sorted; normally, the two highest CC values can be discriminated from the remaining references. Figure 5.9 shows the curves of the CC values of the 4 examples.

5.4.3 Parallel alignment in real space and reconstruction

The reference projections from the 3D lattice were not the projections of the electron density map of the real crystal. It is therefore impossible to align the experimental projections directly in real space. However, the imaging techniques in single particle reconstruction can help to determine the translation information in real space. These rotated experimental projections with known orientations were reconstructed into a map. The map was used to generate the reference projections for aligning the experimental projections in real space (translation only) in
Figure 5.8: The experimental projections and reference projections were compared in Fourier space. Here only the best 4 results are shown. The top row shows the amplitude of the Fourier transformation of the matched reference projections, and the bottom row shows the amplitude of the Fourier transformation of the experimental projections; the amplitude of the Fourier transformation of the reference projections was rotated to match the amplitude of the Fourier transformation of the experimental projections (after rotational alignment). More Bragg spots, especially in the high frequency field, can be found on the amplitude of the Fourier transformation of the reference projections. However, the main spots in the low frequency field, which determine the unit cell, were matched.
5. Preliminary results of phasing crystallographic data

Figure 5.9: The CC value curves in Fourier space. These show the amplitude of the Fourier transform of the 10284 reference projections that were aligned to the amplitude of the Fourier transform of the experimental projections by rotation. This figure gives four examples, and the results were sorted by CC values from high to low. The reference projection with the highest CC value is separated from the remaining CCs.
order to calculate the translation \((x, y)\).

The resolution that can be achieved should be around 4 Å, based on the FSC curve (Figure 5.10). Here, the achieved resolution shows that the two maps reconstructed from the two half-data sets individually are consistent at the 4 Å level. Figure 5.11 shows three different views of the reconstructed map. From the map, it is possible to find the regular alignment, which is a typical characteristic of crystals; if the map is checked locally, a suspected amino acid chain is found. It could be that the map is still not correct, and not the true situation of the peptide crystal, but I believe it is close to it. It can be useful for determining the phasing of the diffraction pattern.

![FSC Curve](image)

**Figure 5.10:** The FSC curve of the map reconstructed from the experimental projections for peptide VQIVYK. As the curve shows, the resolution should be around 4 Å.

### 5.5 Conclusion

Cryo-EM images of lysozyme and peptide nano-crystals contain the phasing which is necessary to solve the protein structure from its crystal. The development of techniques has broken many bottlenecks in the method of collecting diffraction patterns of protein crystals, so that
5. Preliminary results of phasing crystallographic data

Figure 5.11: The orthogonal views of the reconstructed map of the peptide. The map size is $64 \times 64 \times 64$ pixels, and the pixel size is 1.055 Å/pixel.

The cryo-EM images of protein crystals could be used to phase them. The unit cell parameters determined from the diffraction patterns were useful as the reference for determining the orientation parameters of the cryo-EM images in 3D space. The method we developed can find the shift vector for each experimental projection of a crystal, and also its orientation.

For the lysozyme data, the shift vector and orientation of the experimental projections were determined in Fourier space and real space. It is important to divide the task of determining the exact position into two steps, because this decreases the complexity and improves the precision. First the 2D matches in Fourier space are found, in order to be consistent with the matched reference projections. Then the rotated experimental projections are aligned to the matched reference projections in real space. The 3D orientations (Euler angles) can certainly be obtained directly from the matched reference projections. This method is robust, since the main spots on the amplitude of the Fourier transformation of the images can be matched, which ensures that the orientation is correct. The translation parameters have to be determined in a different fashion. For the lysozyme data, we determined these translation parameters by shifting the measured images, matching them to the known 3D
structure. The experimental projections were used to reconstruct an
electron density map with resolution at 8 Å. The poor resolution of
the reconstructed map was probably due to many factors, such as the
noise from the cryo-EM, the image processing, and the fact that the
orientations in 3D space were concentrated in a small range. With the
further development of cryo-EM, the noise level should be reduced in
future. For the data processing, it is helpful to correct the CTF, and
even the beam-induced motion; a more effective algorithm could be
developed to determine the signal and retain it as much as possible.
It would also be better to collect these data by ensuring a larger tilt range.

For the peptide data, I used the same procedure but a different ref-
ence model. The purpose was to test whether the orientation could be
obtained from a 3D lattice in which a unit cell was replaced by a single
atom. This makes it impossible to calculate the shift vector by means of
direct comparison with reference projections in real space. Fortunately,
I could use the technique in single particle analysis to calculate the shift
vector. The projections from the 3D lattice are different from the pro-
jections of the simulated density map, but the amplitude of their Fourier
transformation is similar. It is feasible to estimate the orientations of the
experimental projections. The reconstructed map could be useful as the
initial model for phasing the diffraction patterns.
5. Preliminary results of phasing crystallographic data
Chapter 6

Discussion, prospects and summary
6. Discussion, prospects and summary

6.1 Discussion

Proteins play an important role in life processes. We need to solve their structures in order to understand their functions. Current methods can solve molecular structures, but not all of them can work on biological molecules, because of radiation damage. Traditionally, X-ray crystallography is the best way to solve the structures at the atomic resolution level, especially for inorganic molecules. First, the molecules need to be grown into crystals. Some proteins can be used in X-ray crystallography because the size of their crystals is large enough. Other proteins that can grow into crystals, either 2D crystals or 3D micro- and nano-crystals, cannot survive in such a strong radiation environment. Compared with X-rays, electrons have two main advantages: their strong interaction with the atoms in proteins and their lower energy deposit make it possible to measure meaningful data from the smaller crystals. Electron microscopy is therefore a powerful technique for studying macromolecular complexes. 2D electron crystallography is now a routine method solving the structures of membrane proteins, but 3D electron crystallography is not yet perfect for solving 3D micro- and nano-crystals, and is still in its infancy.

About 30% of proteins that crystallize do not produce crystals of sufficient size [100; 101]. X-ray free-electron lasers (XFELs) seem to be the only way to solve these smaller 3D crystals [138], but they cannot be used widely around the world because they require thousands to millions of crystals, and the instruments are expensive. For the 3D micro- and nano-crystals, scientists achieved little until 2013, when multiple diffraction patterns were collected from a single lysozyme micro-crystal, which was solved at 2.9 Å resolution [29]. The new method, named MicroED, aims to solve the micro-crystals. Before MicroED, our research group had collected diffraction data from lysozyme nano-crystals [28]. Hence, it was a great breakthrough in 3D electron crystallography. Here I discuss how these data were collected and processed, and how it is possible
to phase diffraction patterns.

6.1.1 Data acquisition for 3D protein crystals

The great progress made in recent decades has boosted the rate of solving the structures of biological molecules. The fundamental problem in protein crystallography is radiation damage to biological sample. The most important event in the history of cryo-EM, is probably the fast-freezing technique, which made it possible to preserve samples in the native state and resist radiation damage. Data can therefore be collected during a limited time in cryo-EM. Electron crystallography was first developed to solve 2D protein crystals, especially membrane proteins. SPA was also developed for individual molecules in solution. Electron crystallography could solve structures at the atomic resolution level, while SPA could recently reconstruct a map at near-atomic resolution level, and is improving towards achieving atomic resolution level. This means that we could solve structures in both crystalline and in non-crystalline samples, but there were still some proteins that can grow into 3D crystals, either micro-size or nano-size, which had not been solved. They still needed more effort to solve their structures. MicroED, which can collect 90 diffraction patterns for a single lysozyme micro-crystal [29], broke the bottleneck. These data were recorded as tilt series of still exposures under extremely low electron dose. Later it became possible to collect data in the form of a movie, as the crystal is continuously rotated by the microscope stage [32]. This progress is based on the new detectors: direct electron detectors, which are able to count electrons directly and record data quickly.

6.1.1.1 Detectors in cryo-EM

The main contribution to the progress in solving biological structures came from the improvements in hardware, including the electron microscope, the detectors, and so on. After this, improvements in the data processing, including new algorithms, was also essential for the problems
caused by defects of the hardware, such as the noise from instruments, low SNR and low contrast in images. Hardware is therefore the most important part in structural biology. As the sample can be preserved in amorphous ice, data acquisition became feasible for biological molecules in cryo-EM, yet it was not good enough to achieve high resolution, because of the radiation damage. In response to this situation, a smart method was developed, which records data from hundreds of crystals: one diffraction pattern or image was recorded from each crystal, and they were then merged together. 2D protein electron crystallography is a good example and worked perfectly for membrane proteins. However, it does not work on 3D protein crystals. 3D electron crystallography requires a set of efficient rotation data from a single crystal, either diffraction patterns or images, to solve their structures. Hence, the problem was how to record data faster during a limited time.

Detectors play the key role in data acquisition, and are the core for recording rotation data from a single crystal. Before digital detectors, photographic films were used to record the diffraction data. Photographic films provide high resolution but are too slow for data acquisition and processing. Scientists therefore switched to digital detector cameras. The first widely used digital detector was the CCD, which is convenient for digital read-out but achieves only a low resolution. In addition, its read-out rate is not fast enough for biological samples, although it is much faster than that of photographic films. Recently, a new detector appeared that can read out data much faster than the CCD and has a higher DQE value. This is the direct electron detector, which is based on CMOS technology. The high DQE results in high SNR. Compared with the CCD, the direct electron detector counts the electrons directly, without the procedure of converting electrons into photons. The direct electron detector can work in the rolling-shutter mode, which means there is no gap between adjacent images. The faster rate of data read-out inspired a new method that can correct the motion of molecules and further improves the SNR of images; it also shows a way to observe the behavior
of molecules in SPA.

6.1.1.2 Data collection in cryo-EM

The images and diffraction patterns of lysozyme nano-crystals used in our research were recorded using direct electron detectors. For the image data, this was the first time that lysozyme nano-crystals were imaged. These images came from different crystals and were recorded at tilt angles between $-45^\circ$ and $45^\circ$ under an electron dose of 5 to 10 $e^-/\AA^2$. The highest resolution is about 3.5 $\AA$ measured through the Fourier transform, and more detail of the crystals was shown, by identifying and averaging within the mosaic blocks. Indeed, the contrast of images can be improved dramatically by averaging all the patches in the same coherent domain, using a classification method similar to that in SPA, and averaging the patches with a method similar to that in 2D crystallography. For the diffraction pattern data, this was the first time rotation data (30-frame rotation series) were collected from a single nano-crystal using the Timepix quantum area detector. There was at most a 0.1 $e^-/\AA^2$ electron dose for each frame, which is a much lower electron dose than that of recording images. The extremely low radiation therefore made it possible to collect more data. Following our research, Shi et al. succeeded in recording 90 frames of diffraction data from a single lysozyme micro-crystal with the same amount of electron dose, as each frames exposure was 10s under 0.01 $e^-/\AA^2$. They believe that cumulative radiation damage above 9 $e^-/\AA^2$ will cause significant error [29]. The reason why they can record more diffraction data is, I think, the crystal size: the larger the crystal size, the longer the survival time. All in all, it can be concluded that rotation data can be recorded from a single crystal, even a nano-crystal, using an extremely low electron dose. This was good news for 3D crystals, whether micro-crystals or nano-crystals.
6. Discussion, prospects and summary

6.1.2 Data processing for 3D protein crystals

For the diffraction patterns of 3D protein crystals, some existing methods can be used, such as ADT [22; 23; 24] and RED [25; 26; 27], which are useful for recording and integrating data. In addition, MOSFLM[147] can be used for data processing if the data are recorded in the rolling-shutter mode, because the method of data collection is similar to that in X-ray crystallography. Therefore, the high quality diffraction data of 3D crystals boosted the rate of solving their structures after the breakthrough in data acquisition, but the phase information was still estimated using molecular replacement. It was sometimes a problem that there was no solved structure of a similar molecule for the method of molecular replacement.

The collected image data of 3D protein crystals shows another way to phase the diffraction patterns. In my thesis, the image data were collected from many lysozyme nano-crystals. The first step is to improve the contrast and SNR of the images. In addition to the radiation damage, the weak phase object approximation is also valid for the 3D crystals, resulting in poor contrast of images. The contrast of crystals can be enhanced by averaging the patches that have the same orientation, but it is possible that the 3D crystals are not perfect, and may be warped, twinned, cracked or contain mosaic blocks. All of these lead to the problem that the orientation is not uniform. These images were therefore processed in the same way as the images of particles in SPA. For each image, a large number of patches were extracted randomly and analyzed by MSA to classify them into different domains, with the aim of averaging a high contrast image in each domain. We indeed obtained the high contrast averaged images, and verified that the orientation is not uniform. However, this method does not always work. For the data of peptide nano-crystals, the high contrast images could not be averaged using the same procedure. It is possible that the lighter molecular weight of the peptide molecule compared with
lysozyme caused the lower contrast. The Bragg spots could still be found after the Fourier transform, therefore the traditional Fourier filter that was used in 2D electron crystallography can also be applied to the images of 3D crystals with little change. The noise fraction was estimated instead of masking spots directly, as a mask is too rough and some signals would probably have been lost. The principle was based on the hypothesis that the noise was added to the signal, which means that their phases are a right angle. Hence, the radial average of the power spectra could be used as the estimation of the noise fraction, for further removal of the noise in Fourier space. After the inverse Fourier transform, a high contrast image was calculated. In addition to the radial average, the mean value, average value, and median value can also be used to estimate the noise fraction. We can change the flow value up and down around them to estimate noise more precisely.

The second step is to reconstruct a density map as the phase source, in which the orientation determination is the core. As the images were single shot images of crystals with unknown crystal symmetry, unknown unit cell dimensions, unknown relative orientations and no significant 3D information, the orientation is random. In addition, the orientation is not uniform in the same image. Hence, the orientations of the high contrast averaged images are totally random. This means that the methods used in 2D crystallography cannot be applied in this case. However, it is possible to reconstruct a map from them using the methods of SPA. All the averaged images were treated as particles of a crystal, and we tried to determine the five parameters ($\alpha$, $\beta$, $\gamma$, $x$ and $y$) for each averaged image; yet there is an importance difference: the unit cell is periodical in 3D space, unlike the particles in SPA. This makes it very difficult to determine the common center for all the averaged images. Currently, the best way is to determine Euler angles in Fourier space, in which all the amplitudes of images are centro-symmetric, and determine the in-plane shift vectors in real space by projection matching. This protocol works in theory, but has errors in practice, because it is affected...
6. Discussion, prospects and summary

by many factors, such as motion caused by radiation damage, amplitude changes and the similar or opposite orientations that can be matched when aligned in Fourier space. All of these can cause bias in orientation, and also lead the wrong way for the in-plane shift vectors, which were estimated and refined iteratively in real space. Therefore, the final reconstructed density map was low resolution and contained some errors.

After investigating the Euler angles of these averaged images, it appears that it would be helpful to record images at large tilt angles. The range of the beta angle is narrow, which affects the resolution in the vertical direction, in a similar way to the cone problem in 2D electron crystallography. In addition, the fact that the images come from about 200 crystals with random orientations increases the difficulty of the common center problem in the reconstruction. These problems would not exist or would be alleviated if there could be improvement of the hardware, such as the detectors and sample holder, and if data could be collected from a single crystal.

6.1.3 The phase problem in 3D electron crystallography

The low resolution map reconstructed from the images of crystals can be used to phase the diffraction patterns. Ryan succeeded in phasing X-ray diffraction patterns using a low resolution electron microscopy map through molecular replacement [99]. Here, the map was reconstructed from images of single particles, not images of crystals, therefore the procedure is different. First it is necessary to generate the structure factors and phases from the map, and determine the center and extent of the density. The map density needs to be placed in a large unit cell. It is also necessary to consider the magnification error, and find the non-crystallographic symmetry. Then the phase needs to be extended. In our case, this could be done in the following steps: projections must be generated from the map and transformed into reciprocal space.
Then the next step is to calculate the CC value between them and the diffraction patterns; once matched, the corresponding phase could be passed to the diffraction patterns. The initial phase probably has errors, but these may be acceptable initially and may be refined. The diffraction pattern processing can refine the phase information further to obtain high resolution structures.

The whole procedure is not straightforward, which increases the possibility of errors. The important factor is the quality of image data. Since the images come from different crystals, the orientations are a major problem. For a single crystal, the recorded images also need to be corrected, for either the motion or CTF. Since three images were collected from a single crystal at the tilt angle range -45° and 45°, the beam-induced charging could cause an additional shift of the crystal, and hence a slight difference between adjacent images and a different orientation. The atoms in the sample could have become charged when the beam hit them, which would cause additional phase shift. All in all, this still needs to be refined in order to improve the quality and resolution of the reconstructed map.

6.2 Prospects

The great progress made in cryo-EM has boosted the rate of solving protein structures. In future, some aspects still need to be improved, from the hardware to the software. Better microscopes, better detectors and better software are needed. All of these must aim to collect enough data with sufficiently high SNR and contrast. When the bottleneck has been broken in the hardware, the software can be refined in algorithms, which will further improve the quality of images, and high resolution structures will be obtained.
6. Discussion, prospects and summary

6.2.1 Detectors

The detectors, especially electron counting detectors, have greatly helped with recording the image data. However, this does not mean that the improvement has ended. For example, the DQE value is a very important parameter of detectors because it affects the SNR of images. The higher the value of DQE, the higher the SNR of images, which is reduced by the radiation damage. Although keeping the samples in amorphous ice can preserve their structures and delay the effect of radiation damage, detectors with a fast rate of data read-out and a high DQE are very valuable. Currently the DQE of DEDs and electron counting detectors has not reached 100%. There is therefore still room for improvement.

6.2.2 Sample holder

Why is the holder so important? Because it is one of the reasons that the sample cannot be rotated up to 90°, and the maximum rotation is normally about 60°. A holey carbon sample support on a copper grid is widely used, but is still not perfect. The geometry of the sample holder is usually a key factor that limits the sample rotation. It is therefore a good idea to develop a new holder using new materials. These could be nano-materials that have better conductivity and are thinner.

6.2.3 Data processing

The image quality recorded in cryo-EM is affected by many factors, and cannot be interpreted directly; these factors include optical distortions, crystal tilt, multiple scattering and beam-induced motion. Most of them can be fixed. For the motion of molecules in SPA, the motion of all the molecules can be corrected by image processing, but the Brownian motion cannot be avoided. For the rotation data from a single crystal, although scientists can use them to solve the structure, some aspects still need to be improved. Some intensities in the diffraction patterns need to be corrected because they are incomplete. There is also the
problem of multiple scattering.

For the phase problem in 3D electron crystallography, it is feasible to phase the diffraction patterns from the images or the map reconstructed from them. If the rotation image data can be collected from a single crystal, this will be great progress. I think the first and most important step is to collect image data from a single crystal under an extremely low electron dose. Then it will be possible to process them in the same way as in 2D protein crystallography, because the orientations will be fixed. Here, the existing algorithms probably need to be adjusted because of the differences compared with 2D crystals. It is also possible to obtain phases using molecular replacement, in which a low density map is used as the phase source, as in the case of X-ray diffraction patterns [99]. In fact, the map could also be the template for prediction of the high resolution model for homologous molecules by the program I-TASSER [148].

6.3 Summary

Macromolecules play a key role in life processes. They can be fuel and building materials, and can store energy, protect the cell, support structure, perform transport, provide defense, transmit inherited information, and so on. Structure determines function, therefore we must solve their structures. Based on the structural information, we can understand life processes, and can also design drugs and discover new ways to combat virus infections. The traditional main method for solving biological molecules is diffraction. In recent decades, SPA based on cryo-EM has been developing and continues to be improved. The main advantage of cryo-EM is that the sample does not need to be crystallized, which enormously extends the range of research objects. It also avoids the phase problem in crystallography, and the achieved resolution is currently close to atomic level for some macromolecules. Certainly, radiation damage is still the main problem. Radiation damage
has been alleviated, for example by improving the method of specimen preparation, low-dose radiation and new detectors. Recently, rotation diffraction patterns were collected from a single crystal: 90 diffraction patterns from micro-crystals [29]. Our research group has developed electron nano-diffraction using a dedicated electron diffraction camera based on four Timepix Quad quantum area detectors [141]. We have also collected cryo-EM images of lysozyme nano-crystals. The aim of this thesis is to solve the phase problem for cryo-EM images of nano-crystals. My work has focused on developing a new method for reconstructing the structure of protein/peptide from images of the nano-crystal.

Chapter 2 shows how we can measure images of nano-crystals of lysozyme and improve their SNR through the MSA. Proteins are hard to crystallize, and about 30% of proteins that crystallize cannot produce crystals of sufficient size or quality for X-ray diffraction experiments. We therefore tried to collect projections of nano-crystals instead of diffraction patterns, since the improvement of hardware allows us to collect images. The new detector was used and the electron dose was low, which is important for reducing the damage to the specimen, making it possible to record images of crystals. In addition, cooling the specimen to liquid nitrogen temperatures can preserve it close to the native state and also reduce the radiation damage. Even so, images of lysozyme collected with cryo-EM were devoid of contrast because exposure time was limited by the radiation damage, while the recording was at Scherzer focus. Their Fourier transformations indicated crystalline order with discernible Bragg spots and the resolution we could achieve is 4 Å or better. We found that crystals about 100 nm thick yielded the best resolution. In some cases, splitting of Bragg spots can be observed, which could be caused by the non-uniform ordering of the crystal, or by the Ewald sphere curvature. Spots on the high order Laue zone were also observed. After the Fourier transformation of this image, we could calculate the intensity and phase of these spots. When inspecting the number field of phases, the correlation between adjacent
pixels within a Bragg spot was low, which suggests that there are local differences. These images were processed using a Wiener filter, and it was found that moiré blocks exist, and that orientations in different domains are different. Numerous patches (256×256) were therefore cut from the image and analyzed using MSA in IMAGIC; patches from the image can be classified into different groups and averaged in the same group to enhance SNR.

Chapter 3 describes an image filter that was derived from the Wiener filter. The contrast of images of lysozyme crystals was enhanced using MSA. However, this method did not work on all the images of the peptide nano-crystal. The cryo-EM images of the peptide crystal (VIQVYK) still had too much noise after they were processed by the same protocol, probably because its molecular weight is lighter. In this chapter, a Wiener-type filter was developed and succeeded in denoising these images. It is mainly based on the hypothesis that the noise was added to the signal. Therefore, the fraction of noise in the image can be estimated on the basis of the radial average of the power spectrum. The radial average was calculated from the pixels, leaving out the pixels of spots. The algorithm described in this chapter can also determine the spots and their shapes, which can improve the precision. After the image filter was applied to the images, the contrast was dramatically enhanced and the shape of the crystal could be investigated. However, it is possible that some signals which are covered by noise are lost. For example, the variance of noise is much greater than that of the signal. In any case, it is an excellent filter for the noisy images of crystals.

Chapters 4 and 5 describe a method that is able to determine the five parameters (α, β, γ, x, and y) for the cryo-EM averaged images, and also reconstruct a map that can be used to phase the diffraction patterns. The cryo-EM images of lysozyme and peptide were used to test this method. In addition, the structure of lysozyme was solved from the diffraction patterns, and the model based on the unit cell parameters
of the peptide crystal was used as reference. The cryo-EM images were aligned to the reference in Fourier space and real space in order to determine the Euler angles and in-plane shift vectors respectively. The reason for determining the Euler angles (\(\alpha, \beta\) and \(\gamma\)) is calculated in real space. It is important to divide the task of determining the exact position into two steps, because this decreases the complexity and improves the precision. Comparison in Fourier space is robust, since the main spots on the amplitude of the Fourier transformation of the images can be matched, which can help to determine the correct orientation. For the lysozyme data, the experimental projections were used to reconstruct an electron density map with resolution at 8 Å. For the peptide data, we used the same procedure but the reference was created in a different way. The purpose was to test whether the orientation could be obtained from a 3D lattice in which the density of the unit cell was replaced by a single atom. This makes it impossible to calculate the in-plane shift vectors by aligning them directly to the image of a reference model in real space. However, we can reconstruct the map directly without the in-plane shift vector. Then the projections are re-generated from the map as the reference and the experimental projections can be aligned iteratively until the errors are acceptable. The poor resolution of the reconstructed map was caused by many factors, such as the noise from the cryo-EM and the image processing, but the most important factor is that the orientations were concentrated in a small range due to preferred crystal orientations. It would therefore benefit the achieved resolution if images are collected in a larger tilt angle.
Chapter 7

Samenvatting
Macromoleculen spelen een centrale rol in het levensproces. Ze dienen als brandstof en bouwmateriaal, slaan energie op, beschermen de cel, en ondersteunen de structuur, het transport, verzorgen de afweer en dragen erfelijke informatie over etc. De structuur bepaalt de functie. Daarom moeten de structuren ontraadseld worden. Op basis van deze structurele informatie kunnen we het levensproces begrijpen en kunnen ook medicijnen ontwikkeld worden en nieuwe manieren worden gevonden om virus-infecties te bestrijden. Van oudsher is de belangrijkste methode om biologische moleculen te ontraadselen de diffractiemethode. In de afgelopen decennia is SPA gebaseerd op cryo-EM ontwikkeld, en deze methode wordt nog steeds verbeterd. Het belangrijkste voordeel van cryo-EM is dat het monster niet gekristalliseerd behoeft te worden, hetgeen de omvang aan onderzoeksmogelijkheden enorm vergroot. Ook wordt hiermee het faseprobleem in de kristallografie omzeild, en voor sommige macromoleculen nadert de verkregen resolutie het atoomniveau. Het staat vast dat stralingsschade nog steeds het grootste probleem is. Wel is gebleken dat de stralingsschade beperkt kon worden door bijvoorbeeld de methode voor de preparatie van de monsters te verbeteren, en door gebruik te maken van lage dosis straling en de nieuwe detector. Onlangs lukte het rotatiediffractiepatronen te verkrijgen uit n enkel kristal: 90 diffractiepatronen vanuit microkristallen [29]. Onze onderzoeksgroep ontwikkelde elektron-nano-diffractie door gebruik te maken van de dedicated electron diffractiecamera, gebaseerd op vier timepix-quad quantum area detectors [141]. Ook verkregen we cryo-EM-beelden van het nano-kristal-lysozyme. In dit proefschrift wordt geprobeerd om het faseprobleem van de cryo-EM-beelden van nano-kristallen op te lossen. Mijn werk richt zich op het ontwikkelen van een nieuwe methode om de structuur van protene/peptide uit hun nano-kristal-beelden te reconstrueren.

Hoofdstuk 2 laat zien dat we beelden van nanokristallen van lysozyme kunnen meten en hun SNR kunnen verbeteren door het MSA. Proteinen zijn moeilijk te kristalliseren, en circa 30% van de proteinen die kristalliseren produceren onvoldoende grootte of kwaliteit voor een Rntgendiffractie-experiment. Daarom hebben we geprobeerd om pro-
jecties van nanokristallen te verkrijgen in plaats van diffractiepatronen; de verbetering van de hardware maakte dat mogelijk. We gebruikten de nieuwe detector en hielden de toegepaste elektronendosis laag, wat belangrijk is om schade aan het monster beperkt te houden, en het mogelijk te maken om afbeeldingen van kristallen vast te stellen. Door de monsters te koelen met vloeibaar ethaan kunnen ze dicht bij de natieve toestand gehouden worden en wordt ook de stralingsschade beperkt. Desondanks ontbrak het de beelden van lysozyme die verkregen waren via cryo-EM aan contrast door de stralingsschade, de zwakke fase-approximatie en ook de Scherzer focus zorgde voor een laag contrast. Hun Fouriertransformaties wezen op een kristallijne orde met waarneembare Bragg-spots en de resolutie die we konden bereiken is 4 Å of hoger. We stelden vast dat kristallen die zon 100 nm dik waren de beste resolutie opleverden. In sommige gevallen kon de splitsing van Bragg-spots worden waargenomen, hetgeen veroorzaakt kan zijn door de niet-uniforme ordening van het kristal, of door de Ewald-sphere-kromming. Tevens werden spots op de high order Laue zone waargenomen. Na Fouriertransformatie op dit beeld konden we de intensiteit en de fase van deze spots berekenen. Bij inspectie van het fase-nummer-veld bleek de correlatie tussen aangrenzende pixels binnen de Bragg spot laag, wat suggereert dat er lokale verschillen bestaan. Na verwerking van deze beelden door gebruik te maken van een Wiener filter, bleken moire blocks te bestaan, en orintaties in verschillende domeinen zijn verschillend. Daarom werd er een aantal stukjes (256×256) uitgesneden en werden deze geanalyseerd met behulp van MSA in IMAGIC, stukjes van het beeld kunnen worden ingedeeld in verschillende groepen en gemiddeld in dezelfde groep om SNR te verbeteren.

Hoofdstuk 3 heeft betrekking op een beeldfilter afgeleid van het Wiener filter. Het contrast van beelden van lysozyme crystal werd verbeterd door gebruik te maken van MSA. Maar deze methode werkte niet bij alle beelden van peptide-nanokristal. De cryo-EM-beelden van peptidekristal (VIQVYK) blijven te veel ruis houden nadat hetzelfde protocol erop was toegepast. Waarschijnlijk is dat omdat het moleculaire gewicht lichter is. In dit hoofdstuk is een Wiener-achtig filter ontwikkeld,
7. Samenvatting
dat in staat bleek om de beelden van de ruis te ontdoen. Dat is voornamelijk gebaseerd op de hypothese dat de ruis toegevoegd wordt aan het signaal. Daarom kan een schatting gemaakt worden van de breking van de ruis van het beeld die gebaseerd is op het radiale gemiddelde van het vermogensspectrum. Het radiale gemiddelde werd berekend op basis van de pixels, de pixels van spots buiten beschouwing gelaten. Het algoritme dat in dit hoofdstuk beschreven wordt kan de spots en hun vorm bepalen, waardoor de precisie toeneemt. Nadat het beeldfilter op het beeld is toegepast, werd het contrast enorm verbeterd en kon de vorm van een kristal worden onderzocht. Het blijft echter mogelijk dat sommige signalen onzichtbaar blijven door ruisverlies. Bijvoorbeeld als de variatie van de ruis aanmerkelijk groter is dan die van het signaal. Desondanks blijft dit een uitstekend filter voor de beelden van kristallen die ruis vertonen.

In de hoofdstukken 4 en 5 wordt een methode beschreven die de vijf parameters \((\alpha, \beta, \gamma, x \text{ and } y)\) voor de cryo-EM-gemiddelde beelden kunnen determineren, en om een kaart te reconstrueren die kan worden gebruikt om de diffractiepatronen stapsgewijs in te voeren. De cryo-EM-beelden van lysozyme en peptide zijn gebruikt om deze methode te testen. Bovendien is de structuur van lysozyme afgeleid van de diffractiepatronen en het model gebaseerd op de unit-cell-parameters van het peptidekristal is gebruikt als referentie. De cryo-EM-beelden zijn in lijn gebracht met de referentie in de Fourierruimte en werkelijke ruimte om de Eulerhoeken te bepalen respectievelijk de in-plane verschuiving van de vectoren (Eng.: in-plane shift vectors). Dat de Eulerhoeken \((\alpha, \beta, \text{ and } \gamma)\) in de Fourierruimte worden bepaald, heeft te maken met het feit dat de amplitudes van de Fouriertransformatie van het beeld van een kristal centro-symmetrisch zijn. De in-plane verschuiving van de vector \((x, y)\) wordt in de werkelijke ruimte berekend. Het is van belang om de taak van het bepalen van de exacte positie in twee stappen onder te verdelen, omdat dit de complexiteit vermindert en de precisie verbetert. Vergelijking in de Fourierruimte is stevig, aangezien de belangrijkste spots van de amplitudes in de Fouriertransformatie gematched kunnen worden, wat kan helpen
bij het bepalen van de juiste orintaties. De experimentele projecties reconstrueerden voor de lysozyme-data een elektronendensiteit met een resolutie van 8 Å. Voor de peptide-data gebruikten we dezelfde procedure, maar de referentie werd op een andere manier tot stand gebracht. Het doel was om te testen of de orintatie kon worden verkregen door middel van een 3D-raster waarin de dichtheid van de unit cell vervangen werd door een enkel atoom. Daarom is het onmogelijk om de in-plane verschuivende vectoren (Eng.: in-plane shift vectors) te berekenen door ze direct in lijn te brengen met het beeld van het referentiemodel in de werkelijke ruimte. Wel kunnen we de kaart direct reconstrueren zonder de in-plane verschuivende vector (Eng.: in-plane shift vector). Dan worden de projecties van de kaart gehergenereerd om overeenstemming met de experimentele projecties te bereiken. Deze procedure wordt uitgevoerd totdat de fouten worden aanvaard. De matige resolutie van de gereconstrueerde kaart werd door tal van factoren veroorzaakt, zoals de ruis van de cryo-EM, de verwerking van de beelden, maar het belangrijkste is dat de orintaties geconcentreerd waren in een klein gebied vanwege de preferente kristalorintaties. Een grotere kantelhoek kan derhalve helpen bij het bereiken van een hoge resolutie.
7. Samenvatting
Bash and IMAGIC scripts

YAO-WANG LI

auto2 script

#!/bin/bash

echo "alignment work in Fourier and real space"
for i in 'seq 1 77'
do
   ./align-fourier-space.b $i
   ./align-fourier-space-cc.b $i > tmp.plt
   sed -n 35,10318p tmp.plt >$i-fourier-cc.plt
done

align-fourier-space

align-real-space

#!/bin/csh -f

setenv IMAGIC_BATCH 1

setenv IMAGIC_BATCH 1

echo "! "
echo "! "
echo "! IMAGIC ACCUMULATE FILE "
echo "! "
echo "! "
echo "! "
echo "! IMAGIC program: mralign ---------------------------------------"

/opt/imagic/openmpi/bin/mpiexec -np 12 -x IMAGIC_BATCH
/opt/imagic/align/mralign.e_mpi <<EOF

YES
12
FRESH
ALL_REFERENCES
ALIGNMENT
ROTATE
CCF
anchor-cut-zero-amp anchor-aligned-$1 anchor-cut-zero-amp nhr-best-amp,$1 LOWPASS 0.3 180,180
. Bash and IMAGIC scripts

HIGH
0.0, 0.3
YES
EOF

align-fourier-space-cc.b

#!/bin/csh
setenv IMAGIC BATCH 1
echo "!"
echo "! IMAGIC ACCUMULATE FILE"'
echo "!"
echo "! IMAGIC program: headers -----------------------------"
echo "!"
/opt/imagic/stand/headers.e <<EOF
SORT
CORRELATION
anchor-aligned-$1
EOF

align-real-space

#!/bin/bash
if [-f anchor-amp-best-matched.plt]
then
rm anchor-amp-best-matched.plt
fi
#extract the best hit in fourier space
for i in 'seq 1 77'; do sed -n lp $i-fourer-cc.plt | awk '{print $3}'>> anchor-amp-best-matched.plt;
#extract the amplitude, save it to anchor-amp-best

cat -n anchor-amp-best-matched.plt > tmp
./auto-extract-matched-amp-reference tmp

# parallel alignment
./parallel-alignment-real-space.b

auto-extract-matched-amp-reference

#!/bin/bash
rm anchor-amp-best.img
rn anchor-amp-best.hed
rn anchor-cut-matched-in-fourier-space.hedmn anchor-cut-matched-in-fourier-space.img

#first='sed -n lp $1 | cut -f1'
cat $1 | while read -a line
do
img='echo ${line [0]}'
anchor='echo ${line [1]}'
echo $img
echo $anchor
echo "extracting the matched anchor projections."
./amp-extract.b $img $anchor
./real-extract.b $img $anchor
done

amp-extract.b

#!/bin/csh
setenv IMAGIC BATCH 1
echo "! "
echo "! IMAGIC ACCUMULATE FILE "
echo "! "
echo $1
echo $2
#echo $3
echo "! IMAGIC program: excopy -----------------------------------------------"
echo "$" /opt/imagic/incore/excopy.e <<EOF
2D_IMAGES/SECTIONS EXTRACT
anchor-aligned-$1 anchor-tmp INT
EOF
if (! -e anchor-amp-best.img) then
echo "! "
echo "! "
echo "! IMAGIC program: copyim -----------------------------------------------"
echo "$" /opt/imagic/stand/copyim.e <<EOF
anchor-tmp anchor-amp-best
EOF
else
echo "! "
echo "! IMAGIC program: append -----------------------------------------------"
echo "$" /opt/imagic/stand/append.e <<EOF
anchor-tmp anchor-amp-best
EOF
endif

real-extract.b

#!/bin/csh
setenv IMAGIC BATCH 1
echo "! "
echo "! "
echo "! IMAGIC ACCUMULATE FILE "
echo "! "
echo "$"
echo "$"
#echo "$"
echo "! IMAGIC program: excopy -----------------------------------------------"
Bash and IMAGIC scripts

```
echo "! "
/opt/imagic/incore/excopy.e <<EOF
2D IMAGES/SECTIONS
EXTRACT
anchor-cut
anchor-tmp
INT $2
EOF

if (! -e anchor-cut-matched-in-fourier-space.img) then
  echo "! ________________________________"
  echo "! "
  echo "! "
  echo "! IMAGIC program: copyim -------------------------------"
  echo "! "
  /opt/imagic/stand/copyim.e <<EOF
  anchor-tmp
  anchor-cut-matched-in-fourier-space
  EOF
  else
    echo "! "
    echo "! IMAGIC program: append -------------------------------"
    echo "! "
    /opt/imagic/stand/append.e <<EOF
    anchor-tmp
    anchor-cut-matched-in-fourier-space
    EOF
    endif

parallel-alignment-real-space.b

#!/bin/csh -f
setenv IMAGIC BATCH 1
echo "! "
echo "! ________________________________"
echo "! IMAGIC ACCUMULATE FILE "
echo "! ________________________________"
echo "! "
echo "! "
echo "! IMAGIC program: headers -------------------------------"
echo "! "
/opt/imagic/stand/headers.e <<EOF
PLT
MOVE
anchor-amp-best
anchor-amp-best.plt
EOF
  echo "! "
  echo "! awk '{ print -$2}' anchor-amp-best.plt > rotate-reverse.plt "
  echo "! "
  awk '{ print -$2}' anchor-amp-best.plt > rotate-reverse.plt
  echo "! "
  echo "! IMAGIC program: rotate -------------------------------"
  echo "! "
  /opt/imagic/stand/rotate.e <<EOF
ROTATE
NO
PLT
nhr-best
rotate-reverse.plt
nhr-best-rotate
BILINEAR
NO
EOF
```

echo "!"
echo "! IMAGIC program: alipara
−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
" 
/opt/imagic/align/alipara.e <<EOF
TRANS
CCF
nhr−best−rotate
nhr−best−aligned
anchor−cut−matched−in−fourier−space
0.2
EOF
echo "!"  

echo "! IMAGIC program: headers
−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
" 
/opt/imagic/stand/headers.e <<EOF
PLT
ANGULAR
anchor−cut−matched−in−fourier−space.hed
anchor−cut−matched−eulerangles
EOF
awk '{print $1"\t"$2"\t"$3}' anchor−cut−matched−eulerangles.plt > tmp.plt
echo "!"  
echo "! IMAGIC program: headers
−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
" 
/opt/imagic/stand/headers.e <<EOF
WRITE
ANG−REC
PLT_FILE
tmp.plt
NO
NO
nhr−best−aligned
EOF
. Bash and IMAGIC scripts
References


References


References


References


References


References


References


[98] I. Nederlof, Y.-W. Li, M. van Heel, and J. P. Abrahams, “Imaging protein three-dimensional nanocrystals with cryo-EM,” *Acta crystallo-
References

graphica. Section D, Biological crystallography, vol. 69, pp. 852–9, May 2013. 34, 40, 61, 67, 69, 74, 78, 80


References


152
[129] M. Van Heel, “Classification of very large electron microscopical data sets,” Optik, no. 82, pp. 114–126, 1989. 52


References


References


Curriculum Vitae

Yaowang Li was born on September 1, 1983 in Jiangxi, China. He started his Food Science and Engineer education in Jiangxi Agricultural University in 2004. During the college time, his interest focus on the computer science. In 2008, He obtained a bachelor degree in Food Science.

At the same year, He decided to join the Postgraduate Admission Test, and it was lucky that his Master’s supervisor provide him a project about bioinformatics research at China Agricultural University. Through data analysis with statistical methods, to build the connection between structure and activity for antioxidative peptide.

Shortly after he received a master degree, in 2011, he got the scholarship from China Scholarship Council and joined the project in the department of Biophysical Structural Chemistry, Leiden University, as a PhD candidate. He devoted himself to the research of new methods that can solve the structure of macromolecules base on randomly oriented projections of three-dimensional nano-crystals, which was reported in this thesis.

Currently he will join the Beijing Advanced Innovation Center For Structural Biology at Tsinghua University, continue to the methods research in structural biology.
Publications


